NPRST
Navy Personnel Research, Studies, and Technology
5720 Integrity Drive • Millington, Tennessee 38055-1000 • www.nprst.navy.mil

research at work

NPRST-TR-15-2 February 2015

Technical Guidance for Conducting ASVAB Validation/Standards Studies in the U.S. Navy

Janet D. Held, M.S.
Navy Personnel Research, Studies, and Technology
Thomas R. Carretta, Ph.D.
Air Force Research Laboratory, Wright-Patterson AFB, OH
Sarah A. Hezlett, Ph.D. and Jeff W. Johnson, Ph.D.
Personnel Decisions Research Institute, Inc.
Jorge L. Mendoza, Ph.D.
University of Oklahoma
Norman, M. Abrahams, Ph.D.
Psychometric Solutions
Fritz Drasgow, Ph.D.
University of Illinois at Urbana-Champaign
Rodney A. McCloy, Ph.D.
Human Resources Research Organization
John H. Wolfe, M.A.
Analytical Insight

NPRST-TR-15-2
February 2015

Technical Guidance for Conducting ASVAB Validation/Standards Studies in the U.S. Navy

Janet D. Held, M.S.
Navy Personnel Research, Studies, and Technology
Thomas R. Carretta, Ph.D.
Air Force Research Laboratory, Wright-Patterson AFB, OH
Sarah A. Hezlett, Ph.D. and Jeff W. Johnson, Ph.D.
Personnel Decisions Research Institutes, Inc.
Jorge L. Mendoza, Ph.D.
University of Oklahoma
Norman M. Abrahams, Ph.D.
Psychometric Solutions
Fritz Drasgow, Ph.D.
University of Illinois at Urbana-Champaign
Rodney A. McCloy, Ph.D.
Human Resources Research Organization
John H. Wolfe, M.A.
Analytical Insights

Edited by
Janet D. Held, Thomas R. Carretta, Jeff W. Johnson, and Rodney A. McCloy

Reviewed by
Tanja F. Blackstone, Ph.D.

Approved and released by
David M. Cashbaugh
Director

Navy Personnel Research, Studies, and Technology (NPRST)
Bureau of Naval Personnel (BUPERS-1)
5720 Integrity Drive
Millington, TN 38055-1300
www.nprst.navy.mil

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From-To)
10-02-2015	Technical Report	October 2012-Feb 2014

4. TITLE AND SUBTITLE

Technical Guidance for Conducting ASVAB Validation/Standards Studies in the U.S. Navy

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROJECT ELEMENT NUMBER

6. AUTHORS

Janet D. Held, Thomas R. Carretta, Sarah A. Hezlett, Jeff W. Johnson, Jorge L. Mendoza, Norman M. Abrahams, Fritz Drasgow, Rodney A. McCloy, John H. Wolfe

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATIONS NAME(S) AND ADDRESS(ES)

Navy Personnel Research, Studies, & Technology (NPRST/BUPERS-1)
Bureau of Naval Personnel
5720 Integrity Drive
Millington, TN 38055-1000

8. PERFORMING ORGANIZATION REPORT NUMBER

NPRST-TR-15-2

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The Armed Services Vocational Aptitude Battery (ASVAB) is a joint-service battery used by the U.S. military services for enlistment qualification and occupational classification. All of the military services validate the ASVAB from time to time to ensure that composites of the ASVAB tests they use to classify their enlisted members to occupations are the most predictive of military training performance, and that cutscores are set to manage academically related training attrition. The Navy, however, is the only military service to support an operationally focused ASVAB Validation/Standards program. This technical manual is a synthesis of important technical and procedural information used by the Navy in support of the program and is published so that new industrial/organizational psychologists specializing in tests and measurement within the selection and classification context can have a comprehensive set of guidelines when they are assigned to conduct ASVAB validation/standards studies. The companion manual, Introductory Guide for Conducting ASVAB Validation/Standards Studies in the U.S. Navy, should be read first for that context (NPRST-TR-15-1).

15. SUBJECT TERMS

ASVAB, validity coefficient, restriction in range, reliability, cutscore setting, missing data

16. SECURTIY CLASSIFICATION OF:			17. LIMITATATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
UNCLASSIFIED			UNLIMITED	275	Wendy Douglas
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED			(901)874-2218

Foreword

This is the second of two reports that provide guidance on how to conduct predictive validation studies and set standards for enlisted military occupations using the Armed Services Vocational Aptitude Battery (ASVAB). The ASVAB is the primary enlistment qualification and occupational classification instrument used by all of the U.S. military services. The Navy was the lead on the project because it is the only Service at this time maintaining an operationally focused ASVAB Validation/Standards program. The first report, Introductory Guide for Conducting ASVAB Validation/Standards Studies in the U.S. Navy provides context for the ASVAB and the area of personnel selection and classification whereas this second report provides the technical guidance.

This work was sponsored and funded by the Navy's Selection and Classification office (N132G) with a contribution of funding from the Defense Manpower Data Center – Personnel Testing Division (DMDC - PTD). The work was executed by Navy Personnel Research, Studies, and Technology (NPRST/BUPERS-1), a department of the Bureau of Naval Personnel, along with a team of experts on the various manual topics. The contract work was conducted under the auspices of the U.S. Army Research Office Scientific Services Program administered by Battelle (Delivery Order 0253, Contract No. W911NF-07-D-0001).

David M. Cashbaugh
Director

Dedication/Acknowledgement

This Technical Manual is dedicated to Dr. Nambury S. Raju, whose work in the quantitative area of personnel selection is highly regarded. We sorely missed his planned participation in this project due to his passing.

We would like to acknowledge Dr. Norman M. Abrahams for his tireless and unending contributions, insights, and comments during this manual's conception and writing. We would also like to acknowledge Mr. Robert B. Tiegs, the technical representative for HQ-USMEPCOM on the Manpower Accession Policy Working Group (MAPWG), for his outstanding review that greatly improved both the Introductory and Technical Manuals and, and also, Dr. Jorge L. Mendoza of the University of Oklahoma for his invaluable assistance in interpreting and simplifying difficult psychometric theory.

Executive Summary

The purpose of this Technical Manual, the complement to the Introductory Manual (NPRST-TR-15-1), is to provide background and technical information that will assist those responsible for conducting the studies that result in the setting of military job aptitude/ability standards based on the Armed Services Vocational Aptitude Battery (ASVAB). There are several Department of Defense (DoD) components that have ASVAB responsibilities. The Office of the Under Secretary of Defense for Personnel and Readiness, Accession Policy Directorate, sets policy for the development and use of the ASVAB for determining military service eligibility. The Defense Manpower Data Center - Personnel Testing Division (DMDC - PTD) is the Executive Agent for ASVAB research, development and maintenance. Headquarters, United States Military Entrance Processing Command (HQ-USMEPCOM) is responsible for enlistment processing, which includes maintaining ASVAB testing sites and equipment. Each Service is responsible for developing its own ASVAB job classification composites and cutscores, which we refer to as ASVAB standards. The Manpower Accession Policy Working Group (MAPWG), comprised of technical and policy representatives from the Services, HQ-USMEPCOM, and DMDC, has the responsibility of overseeing the development, effectiveness, and security of the ASVAB, and any new tests that meet the criteria for inclusion in the battery or as adjunct classification tests. Finally, the Defense Advisory Committee on Military Personnel Testing (DACMPT), comprised of nationally recognized experts in the areas of test development and industrial/organizational psychology, provides independent, objective recommendations on ASVAB development and enlistment screening to the Secretary of Defense, through the Under Secretary of Defense for Personnel and Readiness.

The Navy led the development of the two manuals because it is the only Service currently supporting a continuing "ASVAB Validation/Standards Program." All of the Services support ASVAB validation/standards efforts to some degree, but generally (a) on an as-needed basis for specific occupations or occupational groups, (b) periodically when new predictors are considered for occupational classification, or (c) when the validity of the ASVAB is questioned at a highly visible level. The Navy takes the proactive position of conducting ASVAB validation/standards studies on a routine basis because the need is not always apparent. In doing so, the Navy continually monitors potential red flags such as high academically related failure rates or setback rates in training, major changes in the curriculum or training platforms, reductions in training time, recruiting stressors, and the emergence of new occupations (Ratings) or the consolidation of existing Ratings.

Although the Navy follows a general model that addresses ASVAB standards for individual Ratings, ASVAB validation/standards studies are not conducted in a vacuum. A change in an ASVAB standard for one Rating can impact the availability of ASVAB qualified recruits for other Ratings. Rather, the one-Rating study simply means that more individualized attention can be paid to specific issues (e.g., recruiting or training) that influence (moderate) the effectiveness of an ASVAB standard. For example, an individual study conducted for a Rating can result in recommendations beyond the scope of the ASVAB standard, such as (a) allowable ASVAB point waiver maximums, (b) establishment of a course module projected to improve training performance (that

would be much less costly than raising the ASVAB standard so as to severely limit the number of qualified recruits for the Rating), or (c) development of other academic or non-academic screening tools.

As more fully explained in the Introductory Manual, we consider training performance as the criterion upon which to validate the ASVAB because the current version of the ASVAB is comprised of tests with underlying constructs that map well to training curriculum. Another reason training performance rather than job performance is considered the criterion is that the Navy experiences non-trivial academically-related training failures and setbacks at this up front stage. A high rate of training failure translates to high costs for the Navy, but also for the Sailor who might experience a career setback or drop in morale or motivation.

The Introductory Manual provides much more context than this brief introduction and will be of interest not only to the ASVAB validation/standards researcher, but the sponsors of the program and stakeholders (Recruiting, Training, and the Enlisted Community Managers). The contents of the Technical Manual will be of interest mainly to those who actually conduct ASVAB validation/standards studies, or who are in the process of learning how to do so.

The chapters in the Technical Manual are briefly described as follows. Following Chapter 1, the Introduction, Chapter 2 provides a review of basic correlation and regression, classical test theory, and some factors that affect the validity coefficient (validity and correlation are used interchangeably throughout). Chapter 3 provides a discussion about ways to interpret the correlation coefficient. Chapter 4 is about measurement error in our predictor and criterion measures, which can affect the correlation coefficient (validity coefficient). Chapter 5 describes formulas for correcting the validity coefficient for restriction in range. Chapters 6 and 7 are about the joint correction for measurement error and restriction in range. Chapter 8 describes the analytical formulas for deriving the standard error of the range-corrected validity coefficient and cites some of the literature on the bootstrap method. Chapter 9 follows up with a Monte Carlo with bootstrap simulation study that reports on the accuracy and standard errors as a function of a number of study design conditions.

Chapter 10 describes effects on the validity coefficient from violating the assumptions underlying the correction for range restriction. Chapter 11 works through the situation where a negative sign can result for a range-corrected validity coefficient when the sign is positive in the population.

Chapter 12 provides a discussion of commonly applied regression methods and the difficulty in applying statistical power analysis in the context of a restricted in range situation. Chapter 13 is about weighting predictor tests in composites and addresses the tradeoffs between validity and adverse impact. Chapter 14 follows up with more about weighting but from the perspective of a multidimensional performance domain.

Chapter 15 discusses multiple-hurdle selection systems and how bias can be introduced into the estimation of the population validity coefficient when a hurdle is not taken into account. Chapter 16 considers the estimation of the validity coefficient in a multiple-hurdle selection situation within the missing data theory framework.

Chapter 17 is focused on setting cutscores in general, and with application to the ASVAB. Chapter 18 is about simulating recruit job assignments to study the impact of changing an ASVAB standard for one or more Ratings on the fill of all Ratings – a system approach. Chapter 19 is concerned with classification effectiveness and briefly describes previous military models that address differential assignment capability.

The final chapter, Chapter 20, summarizes key points in each chapter of the two manuals attempting to thread together a comprehensive picture of the ASVAB validation/standards methods and concerns.

We advise the reader to be aware that the statistical notation in each chapter follows the preference of the author(s) so is not totally consistent across chapters and sometimes within a chapter, for example, when there are citations of others' work. We note that the reader will often find this situation in the literature so we have not taken the extra time to provide consistency in statistical notation. Statistical terms are defined in each chapter and where not, will be evident.

Finally, we suggest that those in the position to develop ASVAB policy recognize that setting ASVAB standards for military occupations is not a trivial effort and requires not only deep technical knowledge, but a realization that there is always a tradeoff between supply of qualified recruits and the training capacity. Our best hope is that the establishment of a joint-service selection and classification panel (See the charter for the INTERSERVICE Aptitude/Ability Standards Panel in Appendix D of the Introductory Manual) will be proactive in (a) continually monitoring the effectiveness of the ASVAB standards within and across the Services, (b) establishing the integrity of the performance criterion across the Services' schoolhouses, (c) bringing about more hands-on job-like training to the schoolhouses, recognizing that the value of an ASVAB technical test may not fully be appreciated when the criterion measure is strictly academic-based, and (d) exploring use of the most comprehensive and technically sound procedures for conducting predictive validity analyses.

Contents

Figures

Tables

Chapter 1.
Introduction

Purpose/Background

The purpose of this Technical Manual, the complement to the Introductory Manual (NPRST-TR-15-1), is to provide background and technical information that will assist those responsible for conducting the studies that result in the setting of military job aptitude/ability standards based on the Armed Services Vocational Aptitude Battery (ASVAB). The manuals' content, while mainly pertaining to the ASVAB, is broad enough to apply to candidate tests that are yet to be added to the ASVAB, or to be considered as adjunct occupational classification tests. The content also is applicable to those in industry responsible for personnel selection functions, and we draw heavily on non-military research.

As noted in the Introductory Manual, there are several Department of Defense components that have ASVAB responsibilities. The Office of the Under Secretary of Defense for Personnel and Readiness, Accession Policy Directorate, sets policy for the development and use of the ASVAB for determining military service eligibility. The Defense Manpower Data Center, Personnel Testing Division (DMDC-PTD) is the Executive Agent for ASVAB research, development and maintenance. Headquarters, United States Military Entrance Processing Command (HQ-USMEPCOM) is responsible for enlistment processing, which includes maintaining ASVAB testing sites and equipment. Each Service is responsible for developing its own ASVAB job classification composites and cutscores, which we refer to as ASVAB standards. The Manpower Accession Policy Working Group (MAPWG), comprised of technical and policy representatives from the Services, HQ-USMEPCOM, and DMDC as Chairs of the technical committee and full working group, has the responsibility of overseeing the development, effectiveness, and security of the ASVAB, and any new tests that meet the criteria for inclusion in the battery or as adjunct classification tests. Finally, the Defense Advisory Committee on Military Personnel Testing (DACMPT), comprised of nationally recognized experts in the areas of test development and industrial/organizational psychology, provides independent, objective recommendations on ASVAB development and enlistment screening to the Secretary of Defense, through the Under Secretary of Defense for Personnel and Readiness.

At the time of this project the MAPWG was, and still is, fully engaged with the research and processes involved in adding candidate tests to the computer platform that delivers the adaptive version of the ASVAB to military applicants (CAT-ASVAB). These efforts are a result of a commissioned expert ASVAB Review Panel (Drasgow, Embretson, Kyllonen, & Schmitt, 2006). The Panel submitted 21 recommends regarding the ASVAB. One of the recommendations was to validate the use of the ASVAB on a routine basis. The Introductory and Technical Manuals regarding ASVAB validation studies and study methods are meant to fulfill this recommendation.

As stated in the Introductory Manual, a framework or roadmap for conducting ASVAB validation research was developed to address DMDC's goal of having a unified approach that all of the Services could follow (HumRRO) (McCloy, Campbell, Knapp, Strickland, & DiFazio, 2006). The unified framework provides a context for thinking about ASVAB validation research. It outlines diverse validation objectives, reviews different criteria that may be used in validation research, and provides an overview of factors that may influence the Services' capacity to interpret and apply the results of validation studies. The intent of the two manuals' development is to provide more specific information that fulfills these objectives.

The Navy led the development of the two manuals because it is the only Service currently supporting a continuing "ASVAB Validation/Standards Program". All of the Services support ASVAB validation/standards efforts to some degree, but generally (a) on an as-needed basis for specific occupations or occupational groups, (b) periodically when new predictors are considered for occupational classification, or (c) when the validity of the ASVAB is questioned at a highly visible level. The Navy takes the proactive position of conducting ASVAB validation/standards studies on a routine basis because the need is not always apparent. In doing so, the Navy continually monitors potential red flags such as high academically related failure rates or setback rates in training, major changes in the curriculum or training platforms, reductions in training time, recruiting stressors, and the emergence of new occupations (Ratings) or the consolidation of existing Ratings.

Although the Navy follows a general model that addresses ASVAB standards for individual Ratings, ASVAB validation/standards studies are not conducted in a vacuum. A change in an ASVAB standard for one Rating can impact the availability of ASVAB qualified recruits for other Ratings. Rather, the one-Rating study simply means that more individualized attention can be paid to specific issues (e.g., recruiting or training) that influence (moderate) the effectiveness of an ASVAB standard. For example, an individual study conducted for a Rating can result in recommendations beyond the scope of the ASVAB standard, such as (a) tolerable ASVAB point waiver maximums, (b) establishment of a course module projected to improve training performance (that would be much less costly than raising the ASVAB standard so as to severely limit the number of qualified recruits for the Rating), or (c) development new screening tools.

The Navy also conducts occupational group (Rating) studies. Establishing the same ASVAB standard for a homogeneous set or subset of Ratings – that is, with similar levels of training, training time and job complexity – facilitates reassignments of Sailors in the event they are required to cross Ratings because of, say, a military downsizing. Also, having the same ASVAB standard for similar Ratings within an occupational group (if only for a subset of the Ratings) allows the Navy to make initial assignments to the occupational group deferring a specific Rating assignment to a later time when there is more visibility on the Navy's needs (e.g., school seat availability or high losses in the Delayed Entry Program). The final Rating assignment usually occurs upon arrival at Recruit Training Command (RTC), Great Lakes, IL, but could occur later during a core technical course that serves all of the Ratings in the group. Several of the Services follow the occupational group assignment model for at least a portion of their recruiting goals.

As more fully explained in the Introductory Manual, we consider training performance as the criterion upon which to validate the ASVAB because the current version of the ASVAB is comprised of tests with underlying constructs that map well to training curriculum. Another reason training performance rather than job performance is considered the criterion is that the Navy experiences non-trivial academically-related training failures and setbacks at this up front stage. A high rate of training failure translates to high costs for the Navy, but also for the Sailor who might experience a career setback or drop in morale or motivation.

We estimate the ASVAB validity coefficient for the PAY97 normative population (Segall, 2004) rather than yearly Service-specific applicant populations for two reasons. First, validity coefficients can then be compared over time for the same occupation within a Service as part of a monitoring process. Second, validity coefficients can be compared for like occupations across the Services. The across-Service comparison of ASVAB validity coefficients is especially important in this era of downsizing, consolidation of resources, and moves towards conducting joint-service training and operations. Having a common baseline population for validating the ASVAB can help in diagnosing what is accounting for ASVAB validity decay, if it is observed. If ASVAB validity for a particular occupation's training (say, Aviation Mechanic) is observed to be much lower for one Service but not another, the logical question becomes why. That is, is it a training problem or a criterion problem (i.e., inadequate development of performance tests)?

Accurately estimating the population ASVAB validity coefficient is important because cutscores are set in reference to its magnitude. Negative consequences can result from an inappropriately set cutscore, especially when the validity of the ASVAB is high, training is difficult, and there is a substantial performance deficiency. All other things being equal, the larger the validity coefficient the more sensitive the cutscore adjustment will be to improving or degrading future performance levels. When the ASVAB validity coefficient is found to diminish (decay), we should automatically ask why. Many factors affect the validity coefficient; the ones that are statistical or technical are highlighted in the chapters that follow. There are, however, less technical factors that can affect the validity coefficient such as (a) poor training or poor training performance measurement; (b) systematic differences in either the ASVAB testing environment, or the schoolhouse testing environment where training performance is evaluated; and (c) individual differences in motivation. ASVAB examinees should be similarly motivated to test well, and we assume they are when they intend to enlist and qualify for the most desirable military occupations. On the other hand, motivation levels may not be so high when enlisted military members are administered non-ASVAB experimental predictors in the schoolhouses – these students have already passed the operational selection and classification hurdles and are secure in their enlistment decisions as long as they pass the occupational specific training. An unexpectedly low validity coefficient observed for any "experimental" predictor administered in the schoolhouses could be due to unmotivated examinees as well as technical factors. All of these issues are addressed in the chapters that follow.

Besides the Services' role in monitoring ASVAB validity, DMDC constantly monitors the ASVAB for score inflation that could occur due to test compromise – another factor that can result in validity decay. To reduce the possibility of test compromise, DMDC regularly develops new CAT-ASVAB item pools but also applies an item selection algorithm that manages over-exposure of items. We also note that HQ-USMEPCOM has oversight of the administration of the ASVAB across the nation's 65 Military Entrance Processing Stations (MEPS) has been exceptional in ensuring that the ASVAB is administered in standardized secure conditions. Despite all of these efforts to reduce the likelihood of ASVAB compromise (including current efforts to eliminate paper-and-pencil ASVAB – a target for compromise), the Services are concerned that ASVAB validity could decay due in the future if the military budgets continue to decline and there are insufficient funds to resource all of the components developing, maintaining, and operationalizing, and overseeing the ASVAB.

In developing ASVAB validity coefficients, we would hope that the military will continue to develop stellar training and performance criterion measures that clearly map to the skills, and abilities, and knowledge necessary to perform the job and which, in turn, clearly map to the training for the job. Detecting whether this is true requires both a standardized approach to ASVAB validation/standards studies and a common population by which to gauge validity levels. We could take the position that the most current combined Service applicant population should be used rather than the PAY97 population and justify that position by saying that *anything* that alters the ranking of individuals from what would be expected in the population would have an impact on that validity coefficient. Such impact factors include demographic changes that occur in our nation over time, as well as changes in the economic conditions over time. At the time of the manuals' development, military recruitment has benefited from a poor U.S. economy and shortage of private industry jobs. This situation will likely change in the future and if recruiting becomes more difficult, the military will need to adjust either the ASVAB standards, the resources expended for recruiting, or the resources expended for training. DMDC has charge of monitoring applicant ASVAB scores over time and for flags that would indicate a requirement for new norming study (e.g., applicant population characteristics, potential ASVAB additions, score drift or departures from those obtained from national testing programs).

Finally, we suggest that those in the position to develop ASVAB policy recognize that setting ASVAB standards for military occupations is not a trivial effort and requires not only deep technical knowledge, but a realization that there is always a tradeoff between supply of qualified recruits and the training capacity. Our best hope is for the establishment of a joint-service selection and classification working group that will be proactive in (a) continually monitoring the effectiveness of the ASVAB standards for all of the Services, (b) establishing the integrity of the performance criterion across the Services' schoolhouses, (c) bringing about more hands-on job-like training to the schoolhouses, recognizing that the value of an ASVAB technical test may not fully be appreciated when the criterion is strictly academic-based, and (d) exploring use of the most state-of-the-art procedures for conducting validity analyses.

Target Audience

The Introductory Manual that accompanies this Technical Manual is intended to provide broad-based background and procedural information to individuals with diverse backgrounds and responsibilities in sponsoring, overseeing, and policy makers that implement ASVAB standards. The Technical Manual is intended to explain the statistical methods, theory, and formulas to the practitioner who develops ASVAB standards. We have not, however, limited our perspective to the military context and consider the background and interests of the industry practitioner. In addressing both audiences, the technical material in some chapters is presented again in others not to be taken as redundancies, but from different but relevant perspectives and contexts to underscore and reinforce concepts.

We advise the reader to be aware that the statistical notation in each chapter follows the preference of the author(s) so is not totally consistent across chapters and sometimes within a chapter when there are citations of others' work. Statistical terms are defined in each chapter.

Technical Manual Chapters

The chapters in the Technical Manual are briefly described as follows. Following the Introduction, Chapter 2 provides a review of basic correlation and regression, classical test theory, and some factors that affect the validity coefficient (validity and correlation are used interchangeably throughout). Chapter 3 provides a discussion about ways to interpret the correlation coefficient. Chapter 4 is about measurement error in our predictor and criterion measures. Chapter 5 describes formulas for correcting the validity coefficient for restriction in range. Chapters 6 and 7 are about the joint correction for measurement error and restriction in range. Chapter 8 describes the analytical formulas for deriving the standard error of the range-corrected validity coefficient and cites some of the literature on the bootstrap method. Chapter 9 follows up with a Monte Carlo with bootstrap simulation study that reports on the accuracy and standard errors as a function of many study design conditions.

Chapter 10 describes effects on the validity coefficient from violating the assumptions underlying the correction for range restriction. Chapter 11 works through the situation where a negative sign can result for a range-corrected validity coefficient when the sign is positive in the population.

Chapter 12 provides a discussion of commonly applied regression methods and the difficulty in applying statistical power analysis in the context of a restricted in range situation. Chapter 13 is about weighting predictor tests in composites and addresses the tradeoffs between validity and adverse impact. Chapter 14 follows up with more about weighting but from the perspective of a multidimensional performance domain.

Chapter 15 discusses multiple hurdle selection systems and how bias can be introduced into the estimation of the population validity coefficient when a hurdle is not taken into account. Chapter 16 considers the estimation of the validity coefficient in a multiple-hurdle selection situation within the missing data theory framework.

Chapter 17 is focused on setting cutscores in general, and with application to the ASVAB. Chapter 18 is about simulating recruit job assignments to study the impact of changing an ASVAB standard for one or more Ratings on the fill of all Ratings – a system approach. Chapter 19 is concerned with classification effectiveness and previous military models that address differential assignment capability.

The final chapter, Chapter 20, summarizes key points in each chapter of the two manuals attempting to thread together a comprehensive picture of the ASVAB validation/standards methods and concerns.

Appendix A contains an SPSS version of the multivariate range correction provided by the Center for Naval Analyses and related files and instructions. Appendix B contains a generated Taylor-Russell (1939) table that is useful in estimating expected improvements in success rates given ASVAB estimated validity magnitude and other study parameters. Appendix C is a worksheet that uses such tables to estimate classification decision errors.

Acknowledgement of Others' Military Test Validation Work

The authors of the two manuals acknowledge that all of the military personnel research laboratories and their supporting contractors as well as industry and academia have contributed much to the area of personnel selection and classification research. We do not intend that these manuals imply that the Navy has all of the answers for establishing ASVAB standards for military enlisted occupations so we encourage the reader to delve more deeply into the topics that are only briefly discussed in the manuals' chapters. Many of the references in the chapters that follow will lead to important work by all of the Services, such as the work led by Dr. Michael Rumsey and Dr Len White of the Army Research Institute for the Behavioral Sciences (ARI); Dr. Paul Mayberry, Dr. William Sims, Ms. Catherine Hiatt, and Dr. Neil Carey of CNA on behalf of the Marine Corps; Dr. William Alley, Dr. Malcolm Ree, Dr. Melanie Darby, and Mr. Jim Earles of the Air Force Human Resources Laboratory (AFHRL); and Dr. Edward Alf, Dr. Reynaldo Monzon, and Mr. Paul Foley of Navy Personnel Research and Development Center (NPRDC). We recognize the partnerships that the Services have with the Federally Funded Research and Development Centers, and with the many professors in academia, the list being too long to present here but knowable through the references in the manual chapters.

Chapter 1. References

Drasgow, F., Embretson, S. E., Kyllonen, P. C., & Schmitt, N. (2006). *Technical review of the Armed Services Vocational Aptitude Battery (ASVAB)* (FR-06-25). Alexandria, VA: Human Resources Research Organization.

McCloy, R. A, Campbell, J. P., Knapp, D. J., Strickland, W. J., & DiFazio, A. S. (2006). *A framework for conducting validation research with the Armed Services Vocational Aptitude Battery (ASVAB)* (HumRRO Final Report FR-06-15). Alexandria, VA: Human Resources Research Organization.

Segall, D. O. (2004). *Development and evaluation of the 1997 ASVAB score scale* (Technical Report No. 2004-002). Seaside, CA: Defense Manpower Data Center.

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology, 23,* 565-578.

Chapter 2.
Predictor-Criterion Relations:
A Brief Statistical Overview

Jorge L. Mendoza

Introduction

This chapter discusses correlation and regression analyses and the interpretation of these two statistical procedures in the context of personnel selection test validation research. Consider a simple example involving two variables, a single predictor and a criterion. The result of a correlation analysis is an index that depicts the magnitude of the relation between two variables. The result of the regression analysis, simply stated, is the line that best fits a plotted set of data points that represent standing on the two variables. The two variables that apply in personnel selection research are usually a test that is designed to predict performance and a measure of that performance. Although correlation and regression are often discussed separately, these techniques go hand-in-hand, and both are used to explore the relation between two or more variables.

For these procedures to be useful, we must have personnel selection instruments (predictors) and performance outcomes (criteria) that measure the appropriate domains, are free from contamination, and are reliable. Reliability is defined here simply as a psychometric characteristic of a measure that leads to replicable outcomes over many test administrations. We discuss reliability in this chapter from a psychometric perspective as it relates to fundamental test validity analyses. Other chapters in this manual speak more in depth to reliability and to specific topics that negatively influence test validity research results.

From the organization's perspective, the use of aptitude or other cognitive tests in personnel selection is cost-effective only if the performance of personnel selected for jobs is better than what would have been obtained if no selection system were in place. From the individual's perspective, the personnel selection system, to some degree, ensures that they will not be selected for jobs for which there is high potential for failure. The organization's and the individual's perspectives are consistent. There is never a perfect correlation between personnel selection test scores and training or job outcomes, however, so the best we can do is to include tests in our selection systems that have the highest possible correlation, or "validity," in predicting those job outcomes (correlation and validity coefficient are used interchangeably in this chapter).

The chapter topics are as follows: (a) basic correlation and regression analyses, (b) factors that affect correlation and regression, (c) the effect of unreliability on regression and correlation, (d) overestimating the multiple regression coefficient, (e) incremental validity, (f) tests of hypotheses and confidence intervals, (g) range restriction impact on the correlation coefficient, and (h) standard errors of corrected correlations. Some equations and their derivations are presented where appropriate; however, they are considered introductions to topics and so are expanded in subsequent chapters.

Basic Correlation and Regression

According to Rogers and Nicewander (1988), Sir Francis Galton defined the term "regression" in 1885. A decade later, Karl Pearson developed the correlation coefficient. Pearson's correlation coefficient, *r*, is frequently used in the social sciences and other sciences to describe the relation between two variables. The correlation coefficient is also central to many statistical methods (e.g., factor analysis, structural equations, and cluster analysis). Despite its long history, the nuances of the correlation coefficient are not generally well understood (Falk & Well, 1997).The correlation coefficient between two variables x and y, r_{xy}, is defined by the ratio of the covariance between x and y (S_{xy}) to the product of the standard deviations of x and y (S_x and S_y, respectively),

$$r_{xy} = \frac{S_{xy}}{S_x S_y}. \quad (2\text{-}1)$$

The covariance is the numerator in the correlation formula and, as such, it is an unbounded measure of linear association between two variables. The covariance is defined as the sum of the cross-products of centered variables,

$$S_{xy} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{n} \quad (2\text{-}2)$$

where n is the sample size and S denotes the sample (in some later equations, S and s refer to population and sample, respectively, as we will note).

The correlation, on the other hand, is a bounded measure of linear relation ranging from -1 to 1. A correlation of 1 indicates a perfect positive linear relation between two variables, whereas –1 indicates a perfect negative (inverse) linear relation. The correlation is an index of the magnitude of the relation between two variables and how well the data fit a straight line.

Regression, on the other hand, identifies the straight line that fits the data best. A regression line can be described by two values – the intercept, b_0, and slope, b. With a correlation of +1, all of the x and y plotted data points for individuals fall on a straight line with positive slope. The regression line used to predict the y values from the x values is given in most textbooks as

$$\hat{y} = b_0 + b_{y.x}x. \quad (2\text{-}3)$$

It is helpful to write the actual observation y as the sum of two components, y predicted and the error in that prediction (e) so that

$$y = \hat{y} + (y - \hat{y}) = \hat{y} + e. \quad (2\text{-}4)$$

As we see, the observation y is a linear function of two components: the part of y that is predicted by x and the part of y that is not, the residual (error), e. By definition these two components are not correlated, thus allowing the variance of y to be expressed as the sum of the two independent component variances (leaving out the x notation),

$$S_y^2 = S_{\hat{y}}^2 + S_e^2 . \quad (2\text{-}5)$$

Accordingly, the proportion of the total y variance that is predicted from x is given by the ratio

$$\frac{S_{\hat{y}}^2}{S_y^2} . \quad (2\text{-}6)$$

This variance ratio is important in applied settings because it tells us how much of the variability in y can be accounted for by x. If we account for a large proportion of the variability in y, then we know that x is a good predictor of y. For example, if 25% of the variability in school grades is accounted for by the selection test x, then we know that the test score is a relatively good predictor of school performance. We also know that there are other components of performance that are not predicted by the test (75%), which we may be interested in understanding. We can also show that this ratio, the proportion of y variance that is predicted from x, is equal to the correlation coefficient squared (r^2), the square root of which is simply r. (We note that there are a number of ways to derive a Pearson correlation not shown in this chapter, and also a number of coefficients of association).

The intercept of a regression line is often not reported in validation research studies because it (a) does not give us information about the strength of relation between x and y and (b) is tied to the scale of the y variable (which is usually uninformative). Therefore, the focus here is on the slope of the regression line, which is given by the ratio of the covariance to the variance of x:

$$b_{y.x} = \frac{S_{xy}}{S_x^2} . \quad (2\text{-}7)$$

Accordingly, the relation between the correlation and the slope of the regression line is

$$r_{xy} = \frac{b_{y.x} S_x}{S_y} . \quad (2\text{-}8)$$

Note that when x and y are standardized (with mean zero and variance 1), the correlation coefficient equals the regression coefficient. Thus, one interpretation of the correlation coefficient is as the standardized regression coefficient when regressing y on x (Rodgers & Nicewander, 1988). Intuitively, running regression on a correlation matrix instead of a data file with raw scores for each variable produces only the standardized regression coefficient (as opposed to the unstandardized regression weight (slope) and intercept values that apply to raw scores).

We can gain additional insight into the correlation coefficient by rewriting it in terms of *y* residuals and *y* total variance, the proportion of the *y* variance that is *not* predicted from (accounted for) by *x*. The correlation using these terms can be expressed as

$$r_{xy} = \sqrt{1 - \frac{\Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2}} . \quad (2\text{-}9)$$

The numerator under the square root sign in Equation 2-9 indicates the magnitude of the departure of the predicted values from the observed values (residuals) and therefore how well the regression line fits the data. The denominator, which fixes the ratio value between –1 and 1, indicates the magnitude of the departure of the observed values from the mean of those values and is the measure of total *y* variance. If *x* does not predict *y* at all, the numerator and the denominator will be the same value and the ratio will equal 1, with r_{xy} equaling zero. Conversely, if prediction is perfect and all data points are on a straight line, the residuals will equal zero, as will the ratio and therefore r_{xy} will equal 1.

We can gain further insight into the correlation coefficient by squaring Equation 2-9,

$$r_{xy}^2 = 1 - \frac{\Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2} . \quad (2\text{-}10)$$

If we then substitute for the 1 in the equation,

$$r_{xy}^2 = \frac{\Sigma(y - \bar{y})^2}{\Sigma(y - \bar{y})^2} - \frac{\Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2} = \frac{\Sigma(\hat{y} - \bar{y})^2}{\Sigma(y - \bar{y})^2} , \quad (2\text{-}11)$$

we come back to Equation 2-6, the ratio of variance of *y* as predicted by *x* over the variance of *y*:

$$r_{xy}^2 = \frac{S_{\hat{y}}^2}{S_y^2} . \quad (2\text{-}12)$$

It is important to plot the *x*/*y* values when dealing with regression and correlation to inspect linearity and dispersion of scores about the regression line. Figures 2-1 through 2-4 are four scatter plots depicting different values of the correlation coefficient. Figure 2-1 illustrates the absence of relation between *x* and *y* with *a* correlation very near zero (we note scale differences in the x and y axes in the four graphs that should not be a distraction from the intended illustration – the visual forms of the *x*/*y* relationships).

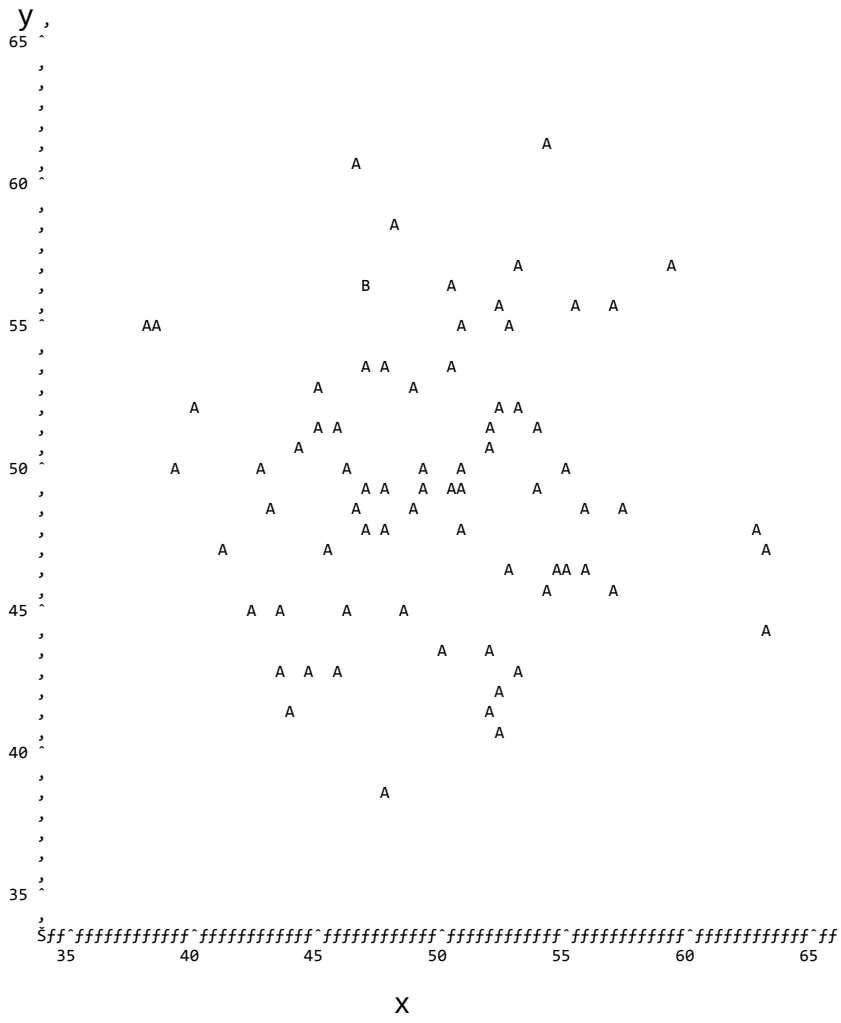


Figure 2-1. Scatter plot of *x* and *y* when r_{xy} = .00.

In Figure 2-2, the correlation is .39.

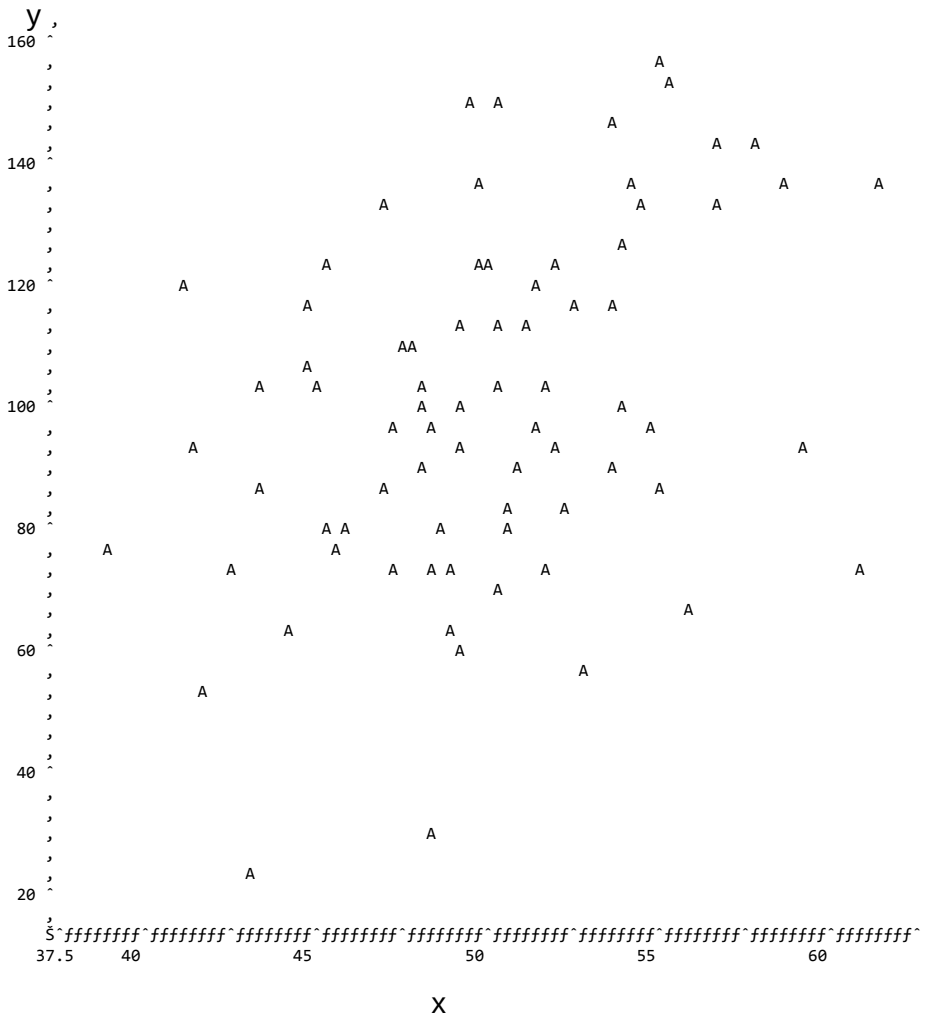


Figure 2-2. Scatter plot of *x* and *y* when r_{xy} = .39.

In Figure 2-3, the correlation is .65

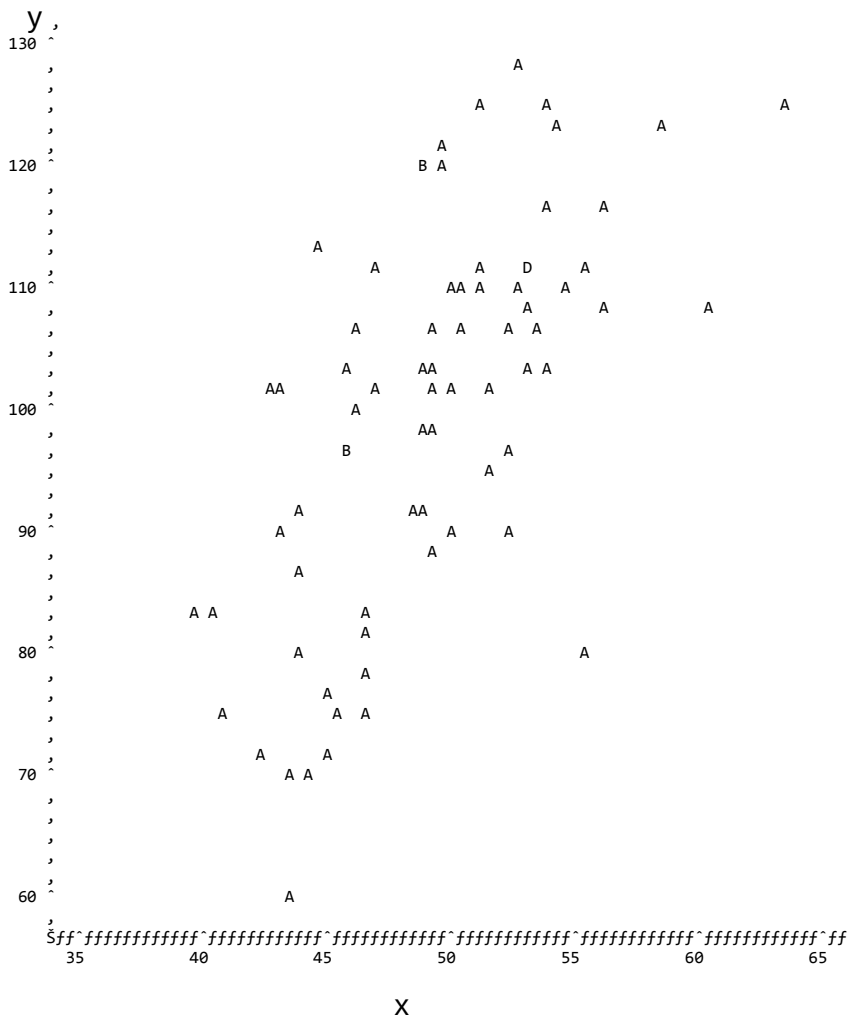


Figure 2-3. Scatter plot of *x* and *y* when r_{xy} = .65.

In Figure 2-4 the correlation is .95.

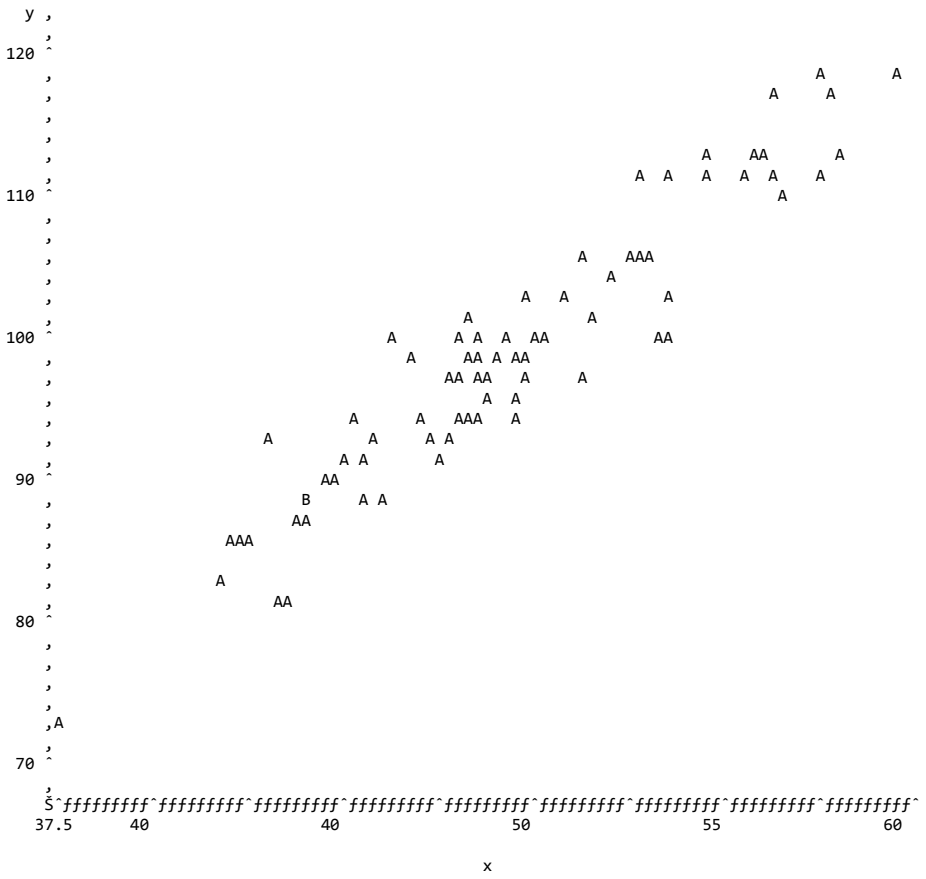


Figure 2-4. Scatter plot of *x* and *y* when r_{xy} = .95.

We can see from Figures 2-1 through 2-4 that as the correlation coefficient approaches 1.00, the scatter plot becomes less circular and more narrowly elliptical implying a stronger relation between *x* and *y*.

Technical Factors That Affect Correlation and Regression

Situational factors can affect the magnitude of the correlation such as, when considering the relation between a personnel selection test and a job performance measure, the motivation of many individuals taking the selection test is low, or as noted in Chapter 1, there is widespread compromise of a high stakes test like the ASVAB. But there are also many technical factors that affect the correlation, six of which we discuss in this section. The first and obvious technical factor that affects the magnitude of the correlation and regression coefficients is the outlier observation. An outlier is a *xy* data point that, graphically in a scatter plot, departs substantially from the observed locations of the rest of the data points.

Outliers are attributed to the response variable (dependent variable) and may or may not have an effect on the regression parameters. That is, an outlier may or may not have influence. For instance, an outlier would not be an influential variable if the bivariate plot showed the data point to be exceedingly far removed from the other data points but running exactly on the regression line. Removing that data point and recalculating the regression line would yield the same slope as when the point was included. Influential variables, on the other hand, may or may not appear as outliers but do affect at least one regression parameter (e.g., slope). Chatterjee and Yilmasz (1992) provided a comprehensive review of the subfield of regression diagnostics that includes a number of useful graphics.

Referencing our past notation on residuals in Equation 2-9, the numerator, we learned, is the sum of the departures of each of the predicted y values from their respective observed y values. Any observation that is sufficiently far away graphically from the rest of the observations could have a large residual. An outlier can either decrease or increase the magnitude of the correlation and regression coefficients. It behooves the researcher to plot the data before conducting a correlation analysis to look for outliers; any suspicious point should be reconsidered carefully. Many standard textbooks on regression and correlation have a section on outliers.

The second technical factor that affects the correlation coefficient and regression coefficients is nonlinearity. Two variables could be related, for example, by a "U" or an inverted "U" function rather than by a straight line. For example, some personality traits may contribute positively to performance up to a certain level, at which point higher levels of the trait become dysfunctional and performance suffers. Fitting a straight line to the scatter plot generated by one of these functions would make little or no sense.

The third technical factor is variability in test scores. If there were little or no variability on either the criterion or predictor, the correlation coefficient would be near zero. We must look for measures that differentiate individuals – that is the major goal of personnel testing research. Measures that provide differentiation in individuals' standing, and the capability of predicting those standings, allows the researcher to apply cutscores that identify the best performers.

The forth technical factor, related to variability in test scores is the restriction in the range of test scores that occurs from applying a cutscore to the selection instrument. Because the correlation coefficient magnitude is associated with test score variance, it will be affected if the full range of variability in test scores is curtailed. The degree of curtailment depends on the stringency of the cutscore, all other things being equal.

The fifth technical factor that affects the correlation coefficient is measurement error, which is a random component of test scores (either the selection instrument – independent variable, or the performance measure – the criterion variable) that affects the precision of a measurement. We do not refer to measurement error as the biased (systematic) contamination of the measurement instrument (e.g., uniform but very dim lighting in an otherwise optimally standardized testing room) but rather the unsystematic "noise." The biasing systematic types of errors are discussed in later chapters. The next section is a discussion of measurement error in the context of classical test theory (CTT).

The sixth technical factor that affects the correlation is the similarity of the two variables' distribution shapes (assuming both are based on a continuous metric scale). The maximum correlation is, of course 1.0, but it has been shown that for large correlations, even moderate departures from distribution similarity (skew and kurtosis) can lower a correlation (not so critical for small correlations, e.g., .30). We refer the reader to Goodwin and Leech (2006) who discussed these six factors impacting the correlation coefficient in greater depth and with excellent references.

Review of the Classical Measurement Model

The classical measurement model (or CTT) defines a test score as the sum of two components, the "true score," t and the error, ε (disregarding sampling error). These components are assumed to be unrelated (i.e., independent) and therefore uncorrelated. We note that CTT does not address sampling error so in using the model we define the criterion y as simply the sum of true ability plus error,

$$y = t_y + \varepsilon_y , \quad (2\text{-}13)$$

and the predictor x as

$$x = t_x + \varepsilon_x . \quad (2\text{-}14)$$

A measure with a small error component (ε) is consistent in measurement over repeated measurements, assuming that the underlying construct (ability or attribute) that we are measuring does not change. The consistency of a measure is defined by the reliability coefficient (not the instrument in and of itself). The reliability coefficient quantifies the "signal-to-noise" ratio of a measure, or "true score variance divided by the total observed score variance."

Given a stable and standardized testing environment, an individual taking the SAT (for example) for college entrance at one setting can expect to obtain nearly the same scores when the SAT is taken again had no additional learning taken place in the interim. The scores will most likely not be exactly the same, however, but affected by what are considered random factors that affect test performance at each testing time such as mood, amount of sleep, worries, or unstable aspects of the test itself that theoretically cancel out (e.g., sometimes a positive mood, sometimes a negative).

There are different ways of measuring test reliability but mathematically, the reliability of x is always given by the ratio of the true score variance of x over the observed score variance,

$$\rho_{xx} = \frac{\sigma^2_{t_x}}{\sigma^2_x} . \quad (2\text{-}15)$$

We can see the similarity of Equation 2-12 and Equation 2-15 and so we can conceptualize the reliability of a test as either (a) the test's correlation with itself (r_{xx}) or (b) the proportion of total test variance that is accounted for by the variance of the true score ($r_{t.x}^2$).

The reliability coefficient ranges from 0 to 1. A reliability of 1 indicates that there is no measurement error; a reliability of 0 indicates that test scores are the result of completely random responses. The test developer needs to know if a potentially useful test is unreliable so that it can be improved. Very often a researcher will discard a test that demonstrates low reliability (or validity) even when there is support from a good theory. Identifying what construct to measure is difficult in and of itself, and when we are successful, it becomes even more difficult to measure it without excessive noise (random factors).

Correcting the Validity Coefficient for Test Unreliability

The following equality illustrates in the theoretical world how to "estimate" the true relation between *x* and *y* correcting for the unreliability of one or the other, or both, measures:

$$r_{t_x,t_y} = \frac{r_{x,t_y}}{\sqrt{\rho_{xx}}} = \frac{r_{t_x,y}}{\sqrt{\rho_{yy}}} = \frac{r_{x,y}}{\sqrt{\rho_{xx}\rho_{yy}}}, \quad (2\text{-}16)$$

where ρ_{xx} is the reliability of *x* and ρ_{yy} is the reliability of *y*. We see from Equation 2-16 that the correlation (validity) is reduced and that we must adjust the observed validity r_{xy} if we are interested in estimating the "true and perfectly reliable" correlation between the two constructs, r_{t_x,t_y}. Note that in the first adjusted correlation in Equation 2-16 we adjust only for criterion unreliability (because the predictor reliability is 1.00, unavailable, or not of interest); whereas in the second adjusted correlation we adjust only for predictor unreliability. Depending upon the purpose of a study (applied or basic construct research), the researcher can correct for unreliability in *x*, *y*, or neither.

The Effect of Unreliability on Regression and Correlation

It is interesting to note that the reliabilities of the *x* and *y* measures set an upper bound to their correlation:

$$r_{xy} \leq \sqrt{\rho_{xx}\rho_{yy}}. \quad (2\text{-}17)$$

For example, assume *x* has a reliability of .90 and *y* has a reliability of .60. If both *x* and *y* were to be corrected for attenuation due to unreliability, the maximum possible correlation that we can observe between these two variables .735 (i.e., 949 times .775). Because the reliabilities determine the upper bound of the criterion-related validity, it is

important to use the most reliable measures possible in both applied and theory-based research.

Another consequence of unreliability is the effect on the regression coefficient – a reduced slope. We can demonstrate (e.g., McNemar, 1962) that the regression coefficient is affected by reliability according to the following equality:

$$b_{t_y.t_x} = b_{y.t_x} = \frac{b_{t_y.x}}{\rho_{xx}} = \frac{b_{y.x}}{\rho_{xx}}. \quad (2\text{-}18)$$

Equation 2-18 tells us that errors of measurement on the predictor affect the regression coefficient (slope), but that errors of measurement on the criterion (not shown) do not. The reason why errors of measurement on the criterion do not affect the regression coefficient is that they end up being part of the residuals, affecting the residual variance but not the slope. As the formulas demonstrate, the regression coefficient does not change with unreliability on y. The standard error (a measure of precision) of b, however, is affected by unreliability (decreasing the standard error with increasing reliability). We can see that the observed regression coefficient $b_{y.x}$ will be less than or equal to the regression coefficient for predicting t_y from t_x by rearrangement of the relevant terms in Equation 2-18:

$$b_{y.x} = \rho_{xx} b_{t_y.t_x}.$$

We note the conceptual parallel of this attenuation formula for the regression coefficient with the attenuation formula for the correlation (e.g., rearranging the relevant terms in the disattenuation formulas in Equation 2-16.

Table 2-1 shows how variability of b systematically differs across the different conditions involving measurement error.

Table 2-1
The Variability of the Slope

Variance of *b*			
$b_{y.x}$	$b_{y.t_x}$	$b_{t_y.x}$	$b_{t_y.t_x}$
$\frac{\sigma_y^2(1-r_{yx}^2)}{n\sigma_x^2}$	$\frac{\sigma_y^2(1-r_{y,t_x}^2)}{n\sigma_{t_x}^2}$	$\frac{\sigma_{ty}^2(1-r_{t_y,x}^2)}{n\sigma_x^2}$	$\frac{\sigma_{ty}^2(1-r_{t_y,t_x}^2)}{n\sigma_{t_x}^2}$

We see from Table 2-1 that the variance of b (typically calculated and referred to as the standard deviation or standard error of b) is the residual variance divided by the

product of the variance of the predictor x and n. Using the results, we can show that $\sigma^2(b_{y.x}) > \sigma^2(b_{t_y.x})$. In other words, errors of measurement in the criterion results in an increase the variability of b.

We close this section by looking at a hypothetical example shown in Table 2-2. Consider a predictor x with reliability .89 and a criterion y with reliability .75. Next, assume that the variance of the predictor is 9 and that the variance of the criterion is 16. The correlation between the predictor x and the criterion y is .42. It is useful to look at the correlation and the slope of the regression line as we define x and y with perfect and imperfect reliability. Table 2-2 gives the correlation and regression coefficients and standard error of b under perfect and imperfect reliability in x and y.

Table 2-2
Correlation and Regression Under Measurement Error

		Predictor					
		x			***t_x***		
		r	*b*	*s(b)*	*r*	*b*	*s(b)*
Criterion	*y*	.416	.555	.1750	.442	.625	.1831
	t_y	.481	.555	.1461	.510	.625	.1520

We can see from Table 2-2 that the correlation is affected by both criterion and predictor measurement error. The regression coefficient, on the other hand, is directly affected only by unreliability in the predictor.

Incremental Validity

Incremental validity (Sechrest, 1963) is the degree to which a measure explains or predicts a phenomenon of interest, relative to other measures. Incremental validity can be evaluated along several dimensions, including the statistical significance of the increase in the correlation coefficient or the increase in the proportion of correct decisions made (e.g., Hunsley & Meyer, 2003). Haynes and Lench (2003) have discussed many indices of incremental validity. In this section, however, we focus on the incremental change in the correlation coefficient.

Consider a situation in which we have two predictors. In this situation the regression would be y on x_1 and x_2; that is,

$$\hat{y} = b_0 + b_{y.x_1}x_1 + b_{y.x_2}x_2 + e. \quad (2\text{-}19)$$

Next assume that x_1 is the "old" predictor and that we want to see the incremental validity of adding the "new" predictor x_2. To do so, we would evaluate the correlation between y and $\hat{y}$ as a function of the two predictors (Equation 2-21) and between y and $\hat{y}$ as a function of one predictor,

$$\hat{y} = b_0 + b_{y.x_1} x_1 + e . \quad (2\text{-}20)$$

The basic idea in evaluating whether x_2 provides incremental validity involves comparing two multiple correlations, $R_{y,x1x2}$ and $R_{y,x1}$.[1] Here, practical as well as statistical issues are important. Because statistical tests of significance vary with sample size, we can find a small difference without practical consequences to be statistically significant. On the other hand, we could miss important differences with a small sample because statistical significance was not attained (not enough power, discussed in a later chapter). Thus, it is good practice to evaluate both the difference in the size of the effect and the statistical significance of the effect. For example, the incremental validity of a candidate test evaluated as a possible addition to the ASVAB could be .04 (observed from the data). However, because the sample size is only 50 cases, that increment might not be detected as statistically significant (recognizing that restriction in range of ASVAB test scores complicates the matter, as discussed in later chapters).

A .04 validity increment to the ASVAB, if real, has practical significance in that a military school with a high academic failure rate would be able to reduce this rate by potentially several percentage points doing nothing else but instating the test. On the other hand, a statistically significant .01 increment in validity found in a sample size of 1,000 would most likely have a trivial impact on graduation rate improvement (possibly 0.5%) but still be worth doing if there are associated improvements in other aspects of the selection and classification system (e.g., reduced adverse impact or increased overall qualification rates across all military occupations for future recruit populations).

Tests of Hypotheses and Confidence Intervals

After showing that two quantitative variables are correlated in a sample, we often want to show that this finding extends to the population. For this purpose, we test the null hypothesis that the population correlation is zero, hoping to reject this hypothesis. Alternatively, we can establish a confidence interval around the population correlation. Under normality, the test that the population correlation is zero is a test of independence. The independence of two variables, assuming normality, involves the following null hypothesis:

$$H_0: \rho = 0.$$

The alternative hypothesis, regardless of whether we are conducting a one-tailed or two-tailed test, is that the variables are related. We either reject or retain this null hypothesis, with its rejection indicating a non-zero relation between the variables in the population. We test this null hypothesis with the t test with n - 2 degrees of freedom. To conduct the statistical test we compare the observed t,

[1] In most situations, it will be multiple correlations that are compared. Here, because of the example involving just two predictors at most, the comparison is between a multiple correlation (when both predictors are used) and a single-order correlation (when just one predictor is used).

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \quad (2\text{-}21)$$

with table values from the *t* distribution that applies to correlation coefficients (found in the appendices of many statistics books). If the observed *t* value that pertains to the sample size falls in the researcher's specified critical upper region of the distribution, then the probability would be too low for the observed sample correlation having occurred by chance and so we reject the null hypothesis. As is the case with all inferential statistics, the *t*-value itself is not the main point of interest. Instead, we are interested in the probability that the *t*-value is large enough to be statistically significant (the *p*-value of the statistic). In many research contexts, a *p*-value smaller than.05 is sufficient to reject the null hypothesis. For example, in Table 2-3 we see that the observed correlation between the *x* and *y* variables for the study sample size of 75 is .31 with an associated *p*-value = .006. The correct decision here is to reject the null in favor of the alternative hypothesis stating that the variables are related in the population.

Table 2-3
Sample Output when Computing Pearson Correlation Coefficients, N = 75

	x	y
x	1.00000	0.31499 (p = 0.0059)
y	0.31499 (p = 0.0059)	1.00000

The *t*-test is somewhat limited, because it cannot be used to test null hypotheses other than $\rho = 0$. Fortunately, however, R. A. Fisher showed how a general test of hypotheses about the population correlation coefficient can be made. It involves transforming the sample correlation into a *z*-score. The Fisher *r* to *z* transformation (discussed in textbooks such as Cohen, Cohen, West, & Aiken, 2003) is given by

$$z(r) = .5\ln\left(\frac{1+r}{1-r}\right), \quad (2\text{-}22)$$

where "ln" stands for the natural log (required because the correlation coefficient distribution is skewed, not symmetrical as with many distributions). For almost any value of ρ, the sampling distribution of $z(r)$ is normally distributed with mean $z(\rho)$ and a large sample variance of $1/(n - 3)$.

This statistical test of a specified magnitude of the correlation is a bit more complicated than the *t*-test for merely the magnitude being greater than zero, because to test the null hypothesis we must transform both *r* and ρ. For example, suppose that instead of stating a null of zero in the previous example, one is interested in testing the null hypothesis that ρ is .50 against ρ is not equal to .50 (either a plus or minus sign). To test the hypothesis, we first transform the null hypothesis in terms of z,

$$H_o: z(\rho = .50) = .55.$$

Next, we transform *r* (= .315) and obtain a $z(r) = .33$, and compute the *z* score

$$z = \frac{.33 - .55}{\sqrt{\frac{1}{72}}} = -1.87.$$

Because -1.87 does not fall in the critical region (< -1.96), we retain the null hypothesis that ρ is .50.

The Fisher *r* to *z* transformation can also be used to construct a confidence interval around the population correlation. The margin of error for a 95% confidence interval on ρ is given by

$$E = \pm z_{1-\alpha}\sqrt{\frac{1}{n-3}} = 1.96(.1178) = .23. \quad (2\text{-}23)$$

so that $z(\rho)$ is .33 ± .23. This confidence interval is too wide to be of practical use. A smaller confidence interval could be obtained by increasing the sample size. One would need a sample size of 200 to bring *E* down to .14, and *n* = 500 to bring it down to .09.

To obtain the confidence interval on ρ instead of $z(\rho)$, we must transform the *z* back to an *r*. We transform it with the following identity:

$$r = \frac{-1 + e^{2z}}{1 + e^{2z}}. \quad (2\text{-}24)$$

The 95% confidence interval for ρ is .315 ± .226.

Confidence Interval on a Predicted Y

Another way of interpreting the correlation coefficient is by considering the error that we make in predicting y from $\hat{y}$. Anytime that we predict an individual y using a regression equation, there is an error associated with the prediction. We can assess the magnitude of this error by constructing a confidence interval around y. Such an interval can be obtained from the t distribution with n = 2 degrees of freedom (*df*), or from many computer programs. Winkler and Hays (1975) showed that this confidence interval is given by

$$\hat{y} \pm t_{(1-\alpha),n-2} \frac{S_y \sqrt{1-r^2}}{\sqrt{n-2}} \sqrt{1+n+\frac{(x-\bar{x})^2}{S_x^2}} . \quad (2\text{-}25)$$

We can see that as r increases, the numerator approaches zero and the interval gets smaller. Consider two values of r, r = .35 versus r = .85. To make matters simple, assume that the variances of y and x are equal to 1, and that n = 100. Also assume that the mean of x is zero and that the value of x is 1. For r = .35, the 95% confidence interval around y is

$$\hat{y} \pm 1.96\left(\frac{.8775}{\sqrt{98}}\right)\sqrt{102} = 1.75$$

When we perform the same calculations for an r of .85, we find

$$\hat{y} \pm 1.96\left(\frac{.2775}{\sqrt{98}}\right)\sqrt{102} = 0.55$$

The width of the interval for the larger r = .85 narrowed to ±0.55 (compared to ±1.75 for the smaller r = .35). We can see that, as the r increases, so does the precision of our predicted value – logically so, because the data points are closer to the regression line.

Figures 2-5 and 2-6 make this point for correlations of .28 and .95, respectively (again noting x/y scale differences should not be a distraction from the visual forms of the graphs).

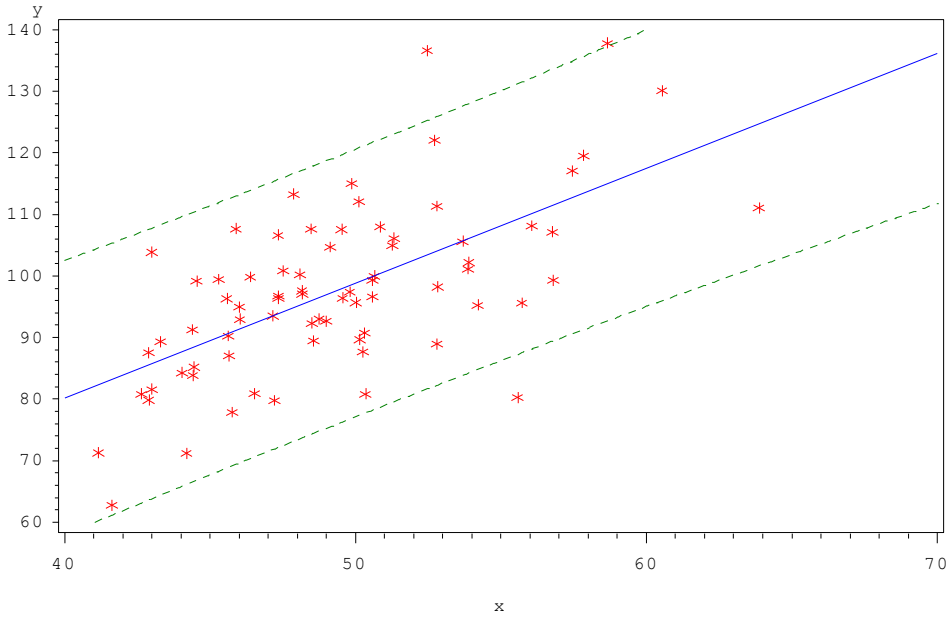


Figure 2-5. 95% confidence interval around *y* when r_{xy} = .28.

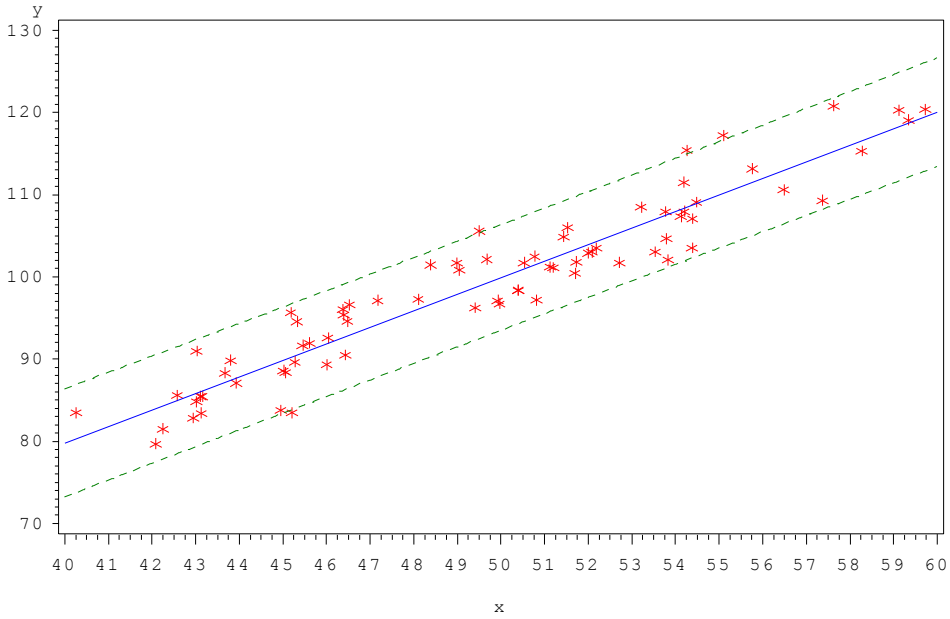


Figure 2-6. 95% confidence interval around *y* when r_{xy} = .95.

In Figures 2-5 and 2-6 we have plotted the 95% confidence interval for each y, obtaining a confidence band around each regression line. As can be seen from these figures, the band around the regression line in Figure 2-5 is much wider than the one in Figure 2-6. The correlation in Figure 2-5 is .28 and the correlation in Figure 2-6 is .95. The larger correlation gives a better estimate of the predicted score, y. Another way to increase the precision of this estimate is by increasing the sample size. Thus, large n's and large correlations are needed when we predict individual scores.

Overestimating the Multiple Regression Coefficient

Multiple regression produces regression coefficients that yield the highest possible multiple *R* in the sample in which they were derived. If these regression coefficients are applied to data in another sample, the *R* will be lower because the regression coefficients are not optimal for that sample. This means the *R* obtained in the sample overestimates the true relation between the set of predictors and the criterion in the population (Cascio, 1991). This phenomenon of over fitting the data, or fitting to the idiosyncrasies of a particular data set, is exaggerated when the sample size is small. There are two procedures commonly used to estimate the over prediction and therefore *R* in the population. The first procedure is cross-validation, in which the optimal predictor weights derived in one (validation) sample are used to predict the criterion in a second (cross-validation) sample. The second procedure involves the application of a formula to estimate the degree of shrinkage in *R* that should occur if in fact we can infer a value for the population.

Cross-Validating with Split Samples

Mosier (1951) provided the classic paradigm for empirical cross-validation in which a single sample is drawn from a population and then divided into a validation and cross-validation sample. Regression weights are estimated in the validation sample and then applied in the cross-validation sample. Murphy (1983) has pointed out that there is only one sampling and that the estimated multiple correlation in the cross-validation sample is still a consequence of "overfitting" the data. Equally important is the fact that, even if there were two samplings from the population, the validation and cross-validation multiple correlations would be only two values out of a virtually infinitely large set of values.

Further, the two-sample cross-validation approach is paradoxical and inconsistent. The goals of estimating regression weights are (a) stability of the estimate and (b) generalizability of the estimated parameters. Weights estimated in two half-samples of n_1 and n_2 cannot be as accurately estimated as from the entire sample of $N = n_1 + n_2$. As is well known, the standard error of a regression weight is a function of the sample size. Splitting a single sample into two pieces reduces the sample size and increases the standard error, thus reducing the accuracy of the estimates (Schmitt, Coyle, & Rauschenberger, 1977). However, the estimation of regression weights in only one sample does not provide estimates of the cross-applicability (i.e., generalizability). The Navy position in validating the ASVAB is that future validations are required involving new samples in the same context in order to "verify" initial findings. Naturally one must take into account not only sampling error, but the characteristics of the sample data that reflect training changes, criterion changes, and demographic changes and that it is entirely possible that an unverified validation reflects real changes.

Estimating Shrinkage with a Formula

A shrinkage formula can be used instead of cross-validation thereby keeping one's sample intact. Use of the formula is less cumbersome and allows the regression coefficients to be estimated in a larger sample (yielding more precise population parameter estimates). There are several "cross-validation" shrinkage formulas (Kennedy, 1988); however, one must consider that they provide estimates that may answer different questions (McCloy, 1994). McCloy points out that the well-known Wherry (1931) shrinkage formula is intended to give an estimate of the multiple correlation in the population. The more relevant question is does the regression equation developed in the sample apply to the population and does the resulting multiple correlation reflect that true but unobservable value.

Cattin (1980) showed that the following formula produces the least biased estimates of the shrunken multiple correlation to be expected if the *sample* equation were applied to the population:

$$\hat{\rho}_c = \sqrt{\frac{(N-k-3)\rho^4 + \rho^2}{(N-2k-2)\rho^2 + k}}, \quad (2\text{-}26)$$

where:
$\hat{\rho}_c$ = estimated population cross-validated multiple correlation,
N = number of people in the sample,
k = number of predictors in the regression equation, and
ρ^2 = population squared multiple correlation

and where ρ must be estimated using the following formula from Wherry (1931):

$$\hat{\rho}^2 = 1 - \frac{N-1}{N-k-1}(1-R^2), \quad (2\text{-}27)$$

where R^2 is the squared multiple correlation in the sample and N and k are as defined above. This value is printed by SPSS in its regression output and labeled "Adjusted R^2." When all predictors are not selected a priori (i.e., the final predictors are selected based on empirical considerations as in stepwise regression), the correction for shrinkage will overestimate $\hat{\rho}_c$ when the final number of predictors is used as the value of k. A more conservative estimate would be obtained by using the original number of predictors (before backward selection) as the value of k which might yield a lower-bound estimate of $\hat{\rho}_c$, and an upper-bound estimate could be obtained by entering the multiple R in the formula for the complete battery of original predictors. The best estimate of $\hat{\rho}_c$ would be the average of the upper- and lower-bound estimates.

The Complication of Range Restriction in Test Scores

When we apply a shrinkage formula to regression results from a sample to estimate the multiple correlation in a population we are not considering the case where a selection instrument with cutscore has been applied to those selected into the sample. Applying a cutscore to a selection instrument produces range restriction in the selected sample's scores because scores are missing below the cutscore. If we restrict in range scores directly/explicitly through the process of selection on *x*, we not only reduce the variability in *x,* but also, indirectly we reduce the variability in *y*. In other words, selection affects not only variances, but also the covariance between *x* and *y*, and thus the magnitude of the correlation coefficient. Further, in theory, when all the assumptions for correcting for range restriction are adhered to (discussed in depth in later chapters), selection affects the correlation coefficient but does not affect the regression coefficient or the error (residual) variance (see Ghiselli, Campbell, & Zedeck, 1981, p. 297).

In theory, the regression coefficients are not affected by restriction in range caused by applying a cutscore to an explicit selection variable because, assuming bivariate normality in the unrestricted population, linearity exists throughout the total predictor score range (an assumption that may or may not be met). We can see how the correlation is attenuated by first rewriting the correlation coefficient squared as

$$r_{xy}^2 = 1 - \frac{S_e^2}{S_y^2}. \quad (2\text{-}28)$$

Because the error variance is unaffected by explicit selection on *x* but the *y* variance decreases, the ratio in Equation 2-28 increases and thus r_{xy}^2 decreases. That the *y* variance decreases by a reduction in *x* variance is simply shown by

$$S_y^2 = b_{y.x}^2 S_x^2 + S_e^2, \quad (2\text{-}29)$$

where $b_{y.x}$ is defined by Equation 2-7. Further, where an estimate of the unrestricted *x* variance is available from the applicant (unrestricted) population; we can compute the ratio of the restricted variance to the unrestricted variance. That ratio can be used in the bivariate explicit selection case to correct the correlation for restriction in range and is (using * to denote statistics computed in the restricted sample)

$$\frac{S_x^{*2}}{S_x^2}. \quad (2\text{-}30)$$

Obviously when hiring using *x*, the criterion data *y* (say performance measures) are only available for those selected. Any correlation computed between *x* and *y*, or between some other potential selection measure *z* and *y* is going to be attenuated by the selection process and these direct and indirect "restricted in range" correlations must be corrected if inferences are to be made about the correlations that applies to the unrestricted applicant

population. As we will see in later chapters, it is the correlation in the unrestricted population that matters in personnel selection and classification. When we correct the correlation coefficient using traditional procedures, the two corrections (direct and indirect) are different. We briefly address the direct restriction in range situation now and discuss both direct and indirect restriction in range in depth in Chapter 5.

To find the relation between the restricted and the unrestricted correlation between x and y due to explicit selection on x, we focus on the restricted variance of y (asterisks indicating restricted statistics),

$$S_y^{*2} = b_{y.x}^2 S_x^{*2} + S_e^2 \ . \quad (2\text{-}31)$$

We can write the restricted variance in terms of the error variance and the slope because they are not affected by selection on x. If we then substitute for the slope and the error variance in Equation 2-31, we obtain

$$S_y^{*2} = \frac{r_{xy}^2 S_y^2}{S_x^2} S_x^{*2} + S_y^{*2}(1 - r_{xy}^{*2}) . \quad (2\text{-}32)$$

With some algebraic manipulation, Equation 2-31 reduces to the following ratio,

$$\frac{r_{xy}^{*2} S_y^{*2}}{S_x^{*2}} = \frac{r_{xy}^2 S_y^2}{S_x^2} . \quad (2\text{-}33)$$

Substituting for the unrestricted variance of y in terms of known quantities, we find that

$$\frac{r_{xy}^{*2} S_y^{*2}}{S_x^{*2}} = \frac{r_{xy}^2 (b_{y.x}^2 S_x^2 + S_e^2)}{S_x^2} . \quad (2\text{-}34)$$

Equation 2-34 can be solved for r_{xy} to obtain an estimate of the unrestricted correlation coefficient under direct range restriction. Note that all the terms can be obtained directly from the data from the restricted sample. The equations that we find in the literature (e.g., Lord & Novick, 1968; Sackett & Yang, 2000) are also solutions for correcting the correlation for range restriction, one for direct restriction range given by

$$r_{xy} = \frac{r_{xy}^* \left(\frac{S_x}{S_x^*} \right)}{\sqrt{1 - r_{xy}^{*2} + r_{xy}^{*2} \left(\frac{S_x^2}{S_x^{*2}} \right)}} . \quad (2\text{-}35)$$

Researchers most frequently use Equation 2-35 to adjust the restricted correlation coefficient for direct range restriction. As an example, assume that the observed correlation in the selected sample between x and y is .21, and that the ratio of variances is 5, corresponding to about 34% selection from the top under a normal distribution. The unrestricted correlation is

$$r_{xy} = \frac{.21\sqrt{5}}{\sqrt{1-.21^2+.21^2*5}} = \frac{.46957}{1.2646} = .37. \quad (2\text{-}36)$$

Table 2-4 is shows the effect of direct range restriction on the correlation coefficient for varying degrees of restriction, designated by the ratio of unrestricted variance, *V(x)*, to restricted variance *V′(x)*, and unrestricted correlation values.

Table 2-4
The Effect of Selection under Direct Range Restriction Assuming Bivariate Normality

Selection Ratio	Restricted Variance	V(x)/V′(x)	Correlation in Restricted Sample (r*)				
			r = .90	r = .70	r = .50	r = .30	r = .10
.05	0.1381	7.24	.2742	.1341	.0795	.0434	.0139
.10	0.1691	5.91	.3297	.1636	.0972	.0531	.0170
.15	0.1949	5.13	.3733	.1876	.1118	.0612	.0196
.20	0.2186	4.57	.4115	.2096	.1252	.0686	.0220
.25	0.2416	4.14	.4464	.2305	.1382	.0758	.0243
.30	0.2645	3.78	.4794	.2510	.1510	.0829	.0266
.35	0.2878	3.47	.5109	.2715	.1639	.0901	.0289
.40	0.3118	3.21	.5413	.2923	.1772	.0976	.0313
.45	0.3369	2.97	.5710	.3136	.1909	.1054	.0338
.50	0.3634	2.75	.6001	.3355	.2053	.1135	.0365
.55	0.3917	2.55	.6288	.3584	.2206	.1223	.0393
.60	0.4223	2.37	.6572	.3824	.2369	.1316	.0424
.65	0.4557	2.19	.6853	.4078	.2544	.1419	.0458
.70	0.4928	2.03	.7132	.4350	.2737	.1532	.0495
.75	0.5347	1.87	.7412	.4642	.2950	.1658	.0537
.80	0.5830	1.72	.7692	.4962	.3190	.1803	.0585
.85	0.6405	1.56	.7976	.5317	.3468	.1974	.0642
.90	0.7121	1.40	.8269	.5723	.3802	.2185	.0714
.95	0.8096	1.24	.8582	.6216	.4235	.2467	.0811
1.00	1.0000	1.00	.9000	.7000	.5000	.3000	.1000

Note. *V(x)/V′(x)* is the ratio of unrestricted to restricted variance.

Table 2-4 shows that as the selection ratio gets smaller (more stringent selection) the observed correlation also gets smaller. For example, when we selected 10% of the subjects from a population with a correlation of .50, the restricted correlation is only .0972. In contrast, if we select 90% instead of 10%, the correlation.50 population correlation is only diminished to .3802.

A similar argument can be made for indirect range restriction where not only is the unrestricted correlation (validity) of interest for the test used in explicit selection (x), but also is of interest for a second test that is correlated to x (say z) but not used in the selection process. Assuming selection on x, the indirect range restriction correction formula is given by documented formulas derived years ago and discussed in a later chapter on corrections for range restriction. Sackett and Yang (2000) provide the formula in a more recent publication:

$$r_{zy} = \frac{r^*_{zy} + r^*_{xy} r^*_{xz} \left(\frac{S_x^2}{S_x^{*2}} - 1 \right)}{\sqrt{\left[1 + r^{*2}_{xy} \left(\frac{S_x^2}{S_x^{*2}} - 1 \right) \right] \left[1 + r^{*2}_{xz} \left(\frac{S_x^2}{S_x^{*2}} - 1 \right) \right]}}. \quad (2\text{-}37)$$

As noted earlier, Chapter 5 provides a more in-depth discussion about the range correction formulas and we also encourage the reader not familiar with this topic to read the various typologies of range restriction discussed by Sackett and Yang (2000).

Standard Errors of Range Corrected Correlations

The usual standard error of the correlation coefficient does not apply to correlations corrected for range restriction (we put aside the issue of unreliable measures for this discussion). Consequently, the usual confidence intervals and tests of hypotheses must be modified. Several researchers (e.g., Bobko & Rieck, 1980; Mendoza, 1993; Raju & Brand, 2003) have proposed large-sample estimators for the standard errors of the corrected coefficients. These formula methods are somewhat cumbersome and hard to compute. Thus, we do not discuss them at length here.

Figures 2-7 through 2-9 show the combined effects of the selection ratio and the correlation coefficient on the standard error of the correction for direct range restriction (see Mendoza, 1993).

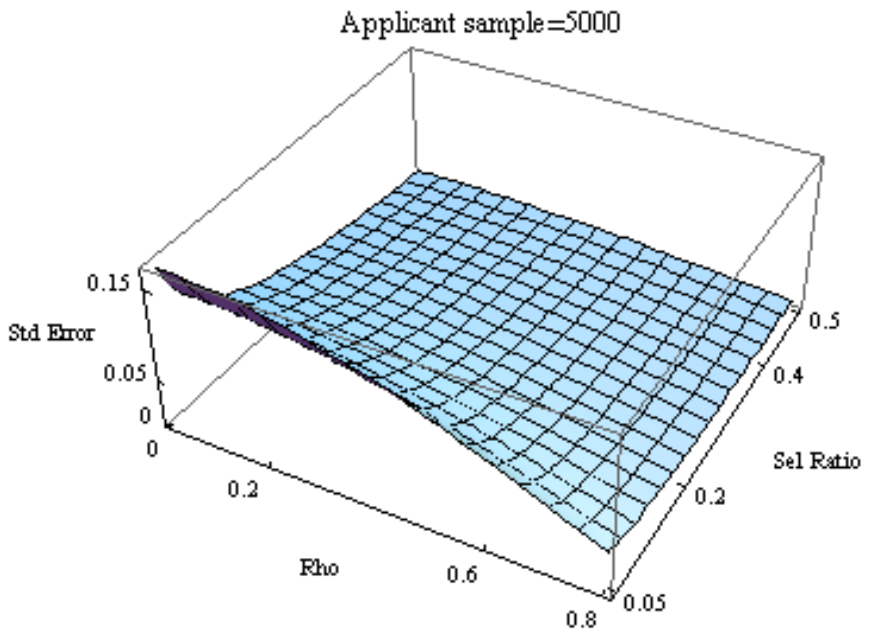


Figure 2-7. Plot of the standard error of the corrected correlation for direct restriction in range when *N* =5000.

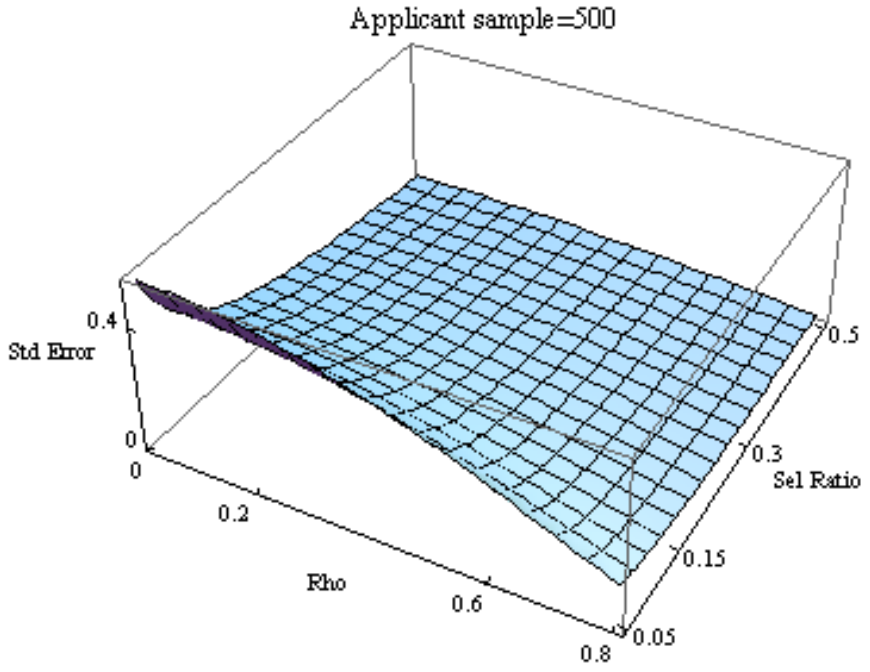


Figure 2-8. Plot of the standard error of the corrected correlation for direct restriction in range when *N* =500.

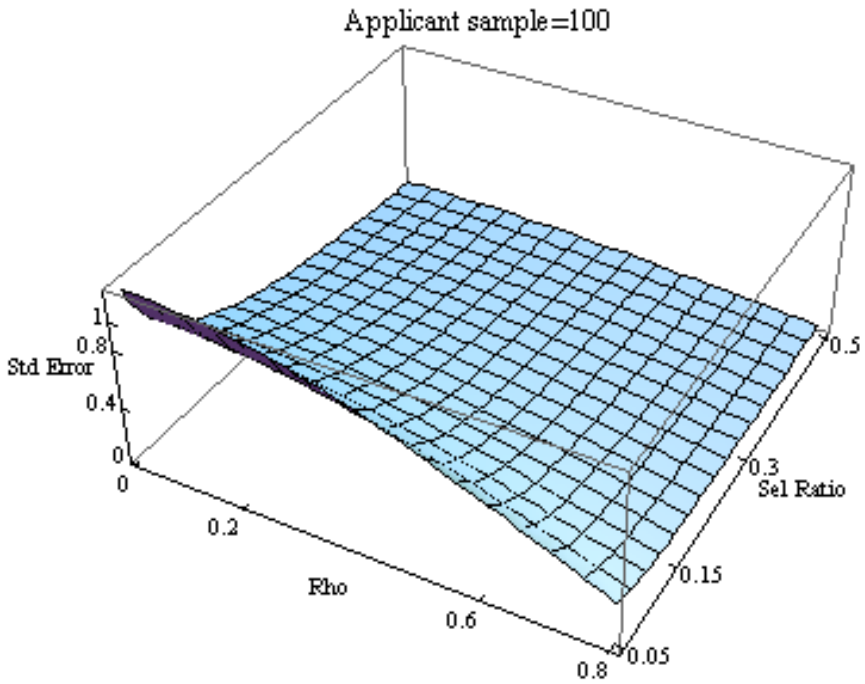


Figure 2-9. Plot of the standard error of the corrected correlation for direct restriction in range when *N* =100.

Figures 2-7 through 2-9 apply to different applicant sample sizes (5,000; 500; and 100, respectively) and their corresponding smaller selected sample sizes at any specific selection ratio. We can see from these figures that the combination of a small applicant sample, a small selection ratio, and a low population correlation results in the largest standard error. This combination of low correlation and a small selected sample size is likely to yield test results with low power and wide confidence intervals.

Although the situation is not much better with computer-intensive procedures, the investigator facing the problem of conducting a hypothesis test or constructing a confidence interval with a corrected estimator should use a computer-intensive procedure—either the bootstrap or multiple imputations. The bootstrap perhaps has more intuitive appeal. The bootstrap has been described fully by Efron and Tibshirani (1993) and for standard error of a corrected correlation by Chan and Chan (2004) and by Mendoza, Hart, and Powell (1991). Although the procedures are slightly different, they give similar results. Under moderate sample sizes with moderate correlations, the bootstrap method has been found to be rather accurate in estimating the standard error under direct range restriction. The case of indirect range restriction has received less attention, but it would appear the bootstrap method should yield similar accuracy in error results. Multiple imputation procedures covering both direct and indirect range restriction also were found to be accurate, as long as the correlation was not close to zero (Chasteen & Mendoza, 2003). More about the bootstrap method is discussed in later chapters.

Concluding Remarks

We have seen that to conduct a high-quality criterion-related predictive validity study we need reliable measures, a quality sample, and personnel selection and performance measures with score variability. Regression analysis was used to identify a linear trend in the data, and the correlation coefficient was used to quantify the fit of the data to the trend. We discussed the importance of creating a scatter plot to examine the data for outliers and nonlinearity, as well as ways to test for statistical significance with one or more predictors. We have seen that the precision of our estimates of *r*, *b*, and *y* depend on the size of both the correlation coefficient and the sample, in addition to other factors that affect correlation and regression analysis. We also reviewed classical test theory, measurement error, confidence intervals, restriction in range of test scores used in selection, the correction for range restriction, and also ways to estimate associated errors in the resulting corrected correlation (validity coefficient), recognizing that sample size is an important factor.

The chapters that follow will provide more in-depth discussion and context on the topics discussed in this chapter. They also will provide greater details about the methods that are most important in accurately estimating the magnitude of the validity coefficient and setting ASVAB standards. The next chapter is a discussion about interpreting the validity coefficient.

Chapter 2. References

Bobko, P., & Rieck, A. (1980). Large sample estimators for standard errors of function of correlation coefficients. *Applied Psychological Measurement*, *4*, 385-398.

Cascio, W. F. (1991). *Applied psychology in personnel management* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

Cattin, P. (1980). Estimating the predictive power of a regression model. *Journal of Applied Psychology*, *65*, 407-414.

Chan, W., & Chan, D. W. L. (2004). Bootstrap standard error and confidence intervals for the correlation corrected for range restriction: A simulation study. *Psychological Methods*, *9*, 369-385.

Chasteen, C. S., & Mendoza, J. L. (2003, April). *Restriction in range issues in validation designs: Modern tools for old problems*. Paper presented at the national meetings of the American Chicago, IL: Educational Research Association.

Chatterjee, S., & Yilmaz, M. (1992). *A review of regression diagnostics for behavioral research*. *Applied Psychological Measurement, 16,* 209-227.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral science: Third Edition*. London: Erlbaum.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. NY: Chapman & Hall.

Falk, R., & Well, A. D. (1997). Many faces of the correlation coefficient. *Journal of Statistics Education*, *5*, 1-14.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. San Francisco: W. H. Freeman and Company.

Goodwin, L. D., & Leech, N. L. (2006). Understanding correlation: Factors that affect the size of *r*. *The Journal of Experimental Education, 74,* 251-266.

Haynes, S. N., & Lench, H. C. (2003). Incremental validity of new clinical assessment measures. *Psychological Assessment*, *15*, 456-466.

Hunsley, J., & Meyer, G. J. (2003). The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. *Psychological Assessment*, *15*, 446-455.

Kennedy, E. (1988). Estimation of the squared cross-validity coefficients in the context of best subtest regression. *Applied Psychological Measurement, 12,* 231-237.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

McCloy, R.A. (1994). Predicting job performance scores for jobs without performance data. In B.F. Green & A.S. Mavor (Eds.), *Modeling cost and performance for military enlistment* (pp. 61-99). Washington, DC: National Academy Press.

McNemar, Q. (1962). *Psychological statistics*. NY: Wiley.

Mendoza, J. L. (1993). Fisher transformations for correlations corrected for selection and missing data. *Psychometrika*, *58*, 601-615.

Mendoza, J. L., Hart, D. E., & Powell, A. (1991). A bootstrap confidence interval based on a correlation corrected for range restriction. *Multivariate Behavioral Research*, *26*, 255-269.

Mendoza, J. L., Stafford, K. L., & Stauffer, J. M. (2000). Large-sample confidence intervals for validity and reliability coefficients. *Psychological Methods*, *5*, 356-369.

Mosier, C. I. (1951). Problems and designs of cross-validation. *Educational and Psychological Measurement, 11,* 5-11.

Murphy, K. R. (1983). Fooling yourself with cross-validation: Single-sample designs. *Personnel Psychology, 36,* 111-118.

Raju, N. S., & Brand, P. A. (2003). Determining the significance of correlations corrected for unreliability and range restriction. *Applied Psychological Measurement, 27,* 52-71.

Rodgers, J. L., & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, *42*, 59-66.

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology*, *85*, 112-118.

Schmitt, N., Coyle, B. W., & Rauschenberger, J. A. (1977). A Monte Carlo evaluation of three formula estimates of cross-validated multiple correlation. *Psychological Bulletin*, *84*, 751-758.

Sechrest, L. (1963). Incremental validity: A recommendation. *Educational and Psychological Measurement*, *23*, 153-158.

Wherry, R. J. (1931). A new formula for predicting the shrinkage of the coefficient of multiple correlation. *Annals of Mathematical Statistics*, *2*, 440-457.

Winkler, R. L., & Hays, L. H. (1975). *Statistics* (2nd ed.). NY: Holt, Rinehart & Winston.

Chapter 3.
Interpreting the Correlation (Validity) Coefficient
Norman M. Abrahams, Jorge L. Mendoza, and
Janet D. Held

Introduction

The ultimate goal of the selection and classification practitioner is to set aptitude/ability standards (e.g., a cutscore applied to a performance predictor) that meet the organization's goal in hiring candidates who perform well enough that the organization succeeds in its mission, given other constraints (training excellence, personnel promotional policies, demand for the product, etc.). The fundamental statistic with which to gage the decision to hire a candidate is the validity coefficient. The previous chapter provided a technical evolution of correlation and regression analysis with a brief review of classical test theory and other statistically based topics. An understanding of these topics is fundamental for an in-depth understanding of the validity coefficient and how it should be interpreted. Unfortunately, although we as personnel research psychologists may fully understand what the validity coefficient means, policy makers may not. This chapter focuses on interpreting the validity coefficient in ways that can better communicate the value of our selection instruments.

Validity Interpretation: An Empirical Expectancy Table

As discussed in Chapter 2, it is often useful to graphically illustrate the relation between two variables. The correlation scatter plots in Figures 2-1 through 2-4 can be modified to provide useful information about the potential for improving the "expected" performance (the criterion – y-axis) by increasing the test score requirement (the predictor – x-axis). In other words, once the correlation between a personnel selection instrument and a performance measure has been determined, the task is to determine how the organization will benefit from the relation. The graphs in Figures 2-1 through 2-4 do not depict the restriction in range that the military observes, but they are useful in explaining "expectancy tables." Given that higher scores on the criterion are associated with higher scores on the predictor, we can collapse the units of each axis of the four figures into fewer but meaningful metric categories. For example, we can collapse the continuous scores on the y-axis into the five grades A, B, C, D, and F to represent grades in college. Also, we can collapse the continuous scores on the x-axis into five percentile categories, or quintiles, to represent standing on the predictor.

There are several variations of the expectancy chart (expectancy table) that apply slightly different quantitative information (Cascio, 1991; Lawshe, 1958). Tables 13 and 14 in the Nuclear Field (NF) study (Appendix B of the Introductory Manual) display the type of empirically based expectancy tables developed for ASVAB standards studies. The ASVAB scores are collapsed on a somewhat arbitrary basis for the NF study and we do so for our example on the next page.

Table 3-1 illustrates an empirically derived expectancy table with an aptitude/ability test used to predict course grades (A, B, C, or D grades). (The illustration uses a very small sample for simplification with the understanding that small sample results are subject to sampling error and are often unstable.)

Table 3-1
Expectancy Table of Grades by Test Scores

	Predictor Test Score									
Grade	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Total**
A	0 0.0%	0 0.0%	0 0.0%	0 0.0%	10 7.1%	30 20.0%	20 20.0%	20 50.0%	20 100%	100
B	0 0.0%	0 0.0%	0 0.0%	20 14.3%	50 35.7%	70 46.7%	50 50.0%	20 50.0%	0 0.0%	210
C	0 0.0%	10 16.7%	50 62.5%	70 50.0%	60 42.9%6	40 26.7%	20 20.0%	0 0.0%	0 0.0%	250
D	20 100%	50 83.3%	30 37.5%	50 35.7%	20 14.3%	10 6.7%	10 10.0%	0 0.0%	0 0.0%	190
Total	**20**	**60**	**80**	**140**	**140**	**150**	**100**	**40**	**20**	**750**

The two variables in Table 3-1 correlate .42 in the observed data so we would expect some systematic differences in the proportion (or percent) of students who attain high and low grades, given their scores on the predictor test. We have rescaled these two continuous variables into a smaller number of meaningful *x* and *y* categories to conveniently illustrate the advantages of an expectancy table. Table 3-1 cell entries contain counts for students who received a particular grade and test score. The sum of each row count is in the far right hand margin, and the sum of each column count is at the bottom margin. The percent entries in each cell are computed by dividing the cell count by its respective column sum; that is, a percentage of individuals achieving a specific grade given a specific predictor score. The probability or expectancy of achieving a specific grade or higher *at a specific test score* is simply the sum of the cell percentages in that test score column for that specific grade and higher.

As an example, we see from Table 3-1 that as the test score increases, the probability (percentage) of students obtaining a C or better increases. We see that at a test score of 2, only 16.7 percent of the students received a C or better whereas at a test score of 6, 93.3% of the individuals received a C or better. As the probability of success increases, the risk of failure decreases. Individuals with higher test scores tend to do better, but the relation is not perfect because the correlation is not 1. For example, there is a large spread of scores in each grade category for the test score of 6 rather than those scores being concentrated solely in the B or C categories. Note that the trend of better performance with higher predictor scores is more likely to appear unstable (or imperfect) when the sample size is small than when it is large.

To illustrate another potential use of the expectancy table, assume we would like to select only those individuals (who apply to take the course) who have *at least* a 50% chance of obtaining a B or A. For ease of interpretation, we recompiled Table 3-1 as Table 3-2, collapsing the grade categories and associated percentages into two labeled "B or better" and "C or worse."

Table 3-2
Expectancy Table of Collapsed Grades by Test Scores

	Predictor Test Score									
Grade	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Total**
B or better	0 0.0%	0 0.0%	0 0.0%	20 14.3%	60 42.9%	100 66.7%	70 70.0%	40 100%	20 100%	**310**
C or worse	20 100%	60 100%	80 100%	120 85.7%	80 57.1%	50 33.3%	30 30.0%	0 0.0%	0 0.0%	**440**
Total	**20**	**60**	**80**	**140**	**140**	**150**	**100**	**40**	**20**	**750**

As we can see from Table 3-2, we would have to select individuals with a test score of 6 or higher to meet the "50% or higher" criterion. At a score of 5, only 42.9% of the individuals receive a grade of B or better (about a .43 probability).

Tables 3-1 and 3-2 are just one of many formats that portray the relationship between a selection instrument's correlation with the criterion of interest, in this case, a predictor test and actual grades in a course for those who enrolled. We refer the reader to other sources mentioned earlier for illustrations. The main point of developing expectancy tables in the personnel selection/classification realm is that quantitative information about two variables can be organized to display their relationship and thus allow interpretation of how adjusting levels on one variable (always the selection instrument) will impact affect levels on the other (the performance, or criterion variable).

Restriction in Range Effect on Interpreting the Correlation

Expectancy tables, in all formats, give a sense of the strength of the relation between the *xy* variables, or predictor and criterion variable, but only for the data at hand. The data at hand in ASVAB validation/standards studies come from schoolhouses that instruct students who have been screened on, among other variables, an ASVAB standard (composite with cutscore). Chapter 2 briefly discussed restriction in range due to a cutscore on an operational (in place) selection/classification instrument. However, Figures 2-1 to 2-4, which show the closeness of data points to the regression lines for broad ranges of scores, do not portray the restriction in range effect on the correlation.

Figure 3-1 clearly shows the effect of a cutscore on the magnitude of the validity coefficient in the practical situation where applicants are selected for training based on an ASVAB standard.

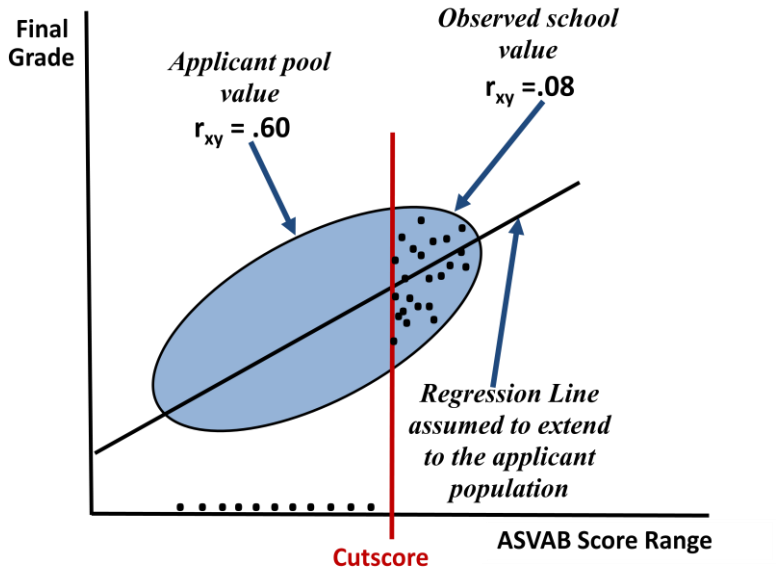


Figure 3-1. Diminished observed validity between ASVAB scores and final school grade when selection is stringent.

Figure 3-1 shows, notionally, a substantial correlation, or validity coefficient, that applies to an applicant population (r_{xy} = .60) if, theoretically, all applicants reflecting a full range of ability (or a random sample of them) were classified to a Rating (occupation) and allowed to attend the Rating's technical training course. Not all applicants are selected, however, as indicated by the dots on the *x*-axis without a partner *y* score. The validity coefficient of the ASVAB measure (a composite of ASVAB tests specific to the Rating) is a much lower r_{xy} = .08 and not of interest. That is, the full-range validity coefficient for the applicant population is of interest because it is *that* validity coefficient that will be used to set an effective cutscore (a balancing act described later). Chapter 2 briefly described a single case of estimating the unrestricted (population) validity coefficient (explicit selection) and Chapter 5 goes further into the topic. The point here is that the restricted validity coefficient is uninterpretable for our process of setting ASVAB standards.

The Taylor-Russell Tables for Interpreting Validity Coefficients

Assume that the correlation in not restricted in range. As noted from long ago (e.g., Hull, 1928), if we then consider the magnitude of the correlation coefficient and its squared value, the amount of shared variance between the test and the performance measure can appear small (e.g., a correlation of .20 squared translates into 4% of two variables' overlap). Taylor and Russell (1939) took a different perspective and showed that even small validity coefficients can be useful. The Taylor and Russell (1939) tables are theory-based and derived from bivariate normal distributions of any positive correlation magnitude (between zero and 1). Figure 3-2 on the next page will be used for illustration of how the tables were compiled and their use in interpreting the utility of a validity coefficient.

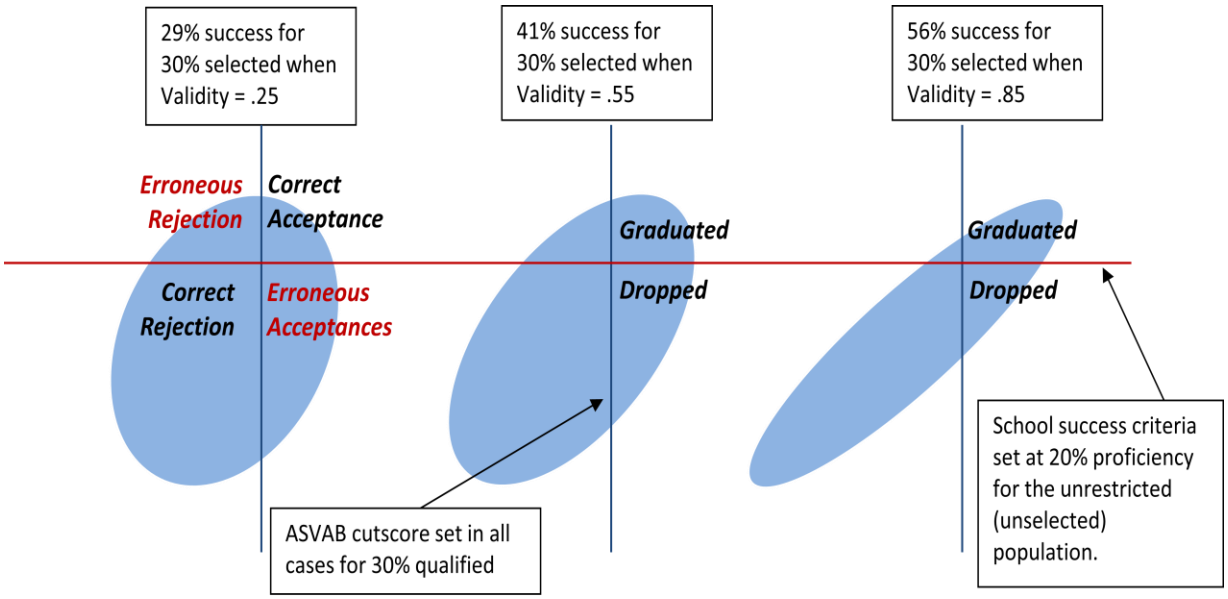


Figure 3-2. Success rate improvement as a function of magnitude of the validity coefficient (base rate = .20).

Figure 3-2 is taken from the Navy's Nuclear Field (NF) ASVAB validation/standards study in Appendix B of the Introductory Manual and is used to illustrate the expected improvement in the graduation rate applying a selection instrument with three validity magnitudes. The three bivariate distributions represent, from left to right, validities of .25 (the smallest ASVAB validity coefficient observed for SEAL physically intensive training criteria), .55 (the average of ASVAB composite validities in predicting training performance across all Navy entry level schools), and .85 (the largest ASVAB validity coefficient observed for academically based Nuclear Field courses).

Overlaid on each bivariate normal in Figure 3-2 is the same *y*-axis performance bar (horizontal) reflecting a 20% success rate (in training) and the same *x*-axis cutscore (vertical) reflecting an ASVAB 30% qualification rate. Both of these "parameters" being equal, the success rate of the selected group is solely dependent on the magnitude of the validity coefficient. The aggregate success rates for the Figure 3-2 graphs are 29% for r_{xy} = .25; 41% for r_{xy} = .55; and 56% for r_{xy} = .85.

The Taylor-Russell (1939) tables are mathematically derived and displayed as 10 published tables each with a different "base rate" success rate, but with the same validity coefficient range (0 to 1.0 displayed in the left most column) and the same selection ratio (SR) range (display across the first header row). The table entries are the success rates associated with each SR and validity coefficient combination, which differ across tables. Also displayed on the left most graphs are the correct and incorrect classification decision outcomes associated with each combination of base rate, SR, validity coefficient magnitude, and success rate, explained shortly.

In the military application of the Taylor Russell (1939) tables, three of the four parameters are known from ASVAB validation/standards studies, so the fourth (the base rate) is fixed. Specifically known are (a) the correlation between the ASVAB composite scores and training grades for the unrestricted ASVAB population is estimated for the unrestricted population applying the correction for range restriction to the correlation estimated in the ASVAB restricted school sample; (b) the selection ratio is the proportion of the applicant population qualified for an occupation due to the operational ASVAB composite's cutscores (the ASVAB normative population is used in most cases for uniformity across studies and because of the full range of ASVAB scores), and (c) the success rate is the pass rate that applies to the school sample.

Finding the applicable Taylor-Russell (1939) base rate table is a simple process of fitting these three parameters to every base rate until a match is found. The match criterion is that the internal table success rate that exists at the intersection of the validity coefficient and the selection ratio matches the school sample success rate. Once the correct base rate table is identified, various assessment scenarios can be evaluated such as (a) estimating improved success rates from raising the cutscore on the operational ASVAB composite, (b) replacing the operational composite with one that is evaluated as having a larger validity coefficient, or (c) lowering the selection ratio (cutscore) if recruitment is a problem and the status quo success rate is not an issue.

Figure 3-2 shows that, all other things being equal (i.e., a cutscore that qualifies 30% of the population, and challenging training that only 20% of applicants would be expected to pass), the validity magnitude being the sole determinant of the selected candidates' success rate (29% success for r_{xy} = .25; 41% for r_{xy} = .55; and 56% for an validity coefficient = .85). A generated Taylor-Russell (1939) table for a base rate of .10 is provided in Appendix A as an example.

The far left graph in Figure 3-2 also depicts another validity interpretation – all other things being equal, the larger validity coefficient (with exceptions), the more accurate the selection/classification decision. As depicted, there are two correct selection decisions (correctly accepted and correctly rejected) and two incorrect selection decisions (incorrectly accepted and incorrectly rejected). One can visualize moving the vertical cutscore bar to the right in all three graphs to improve the success rate and correct selection decisions (both rejecting as well as accepting applicants) but at the expense of increasing the incorrect decision of rejecting applicants who would have succeeded. At least for the military, it may be that the applicant population propensed to enlist in the military diminishes to such an extent that the cost of rejecting able and willing youth becomes larger than accepting them with higher risk.

Appendix B provides a worksheet example of how to compute the correct and incorrect decision quadrants of a bivariate distribution when there is a specified level of performance on both the selection instrument and the criterion. We note at this point that the Taylor-Russell (1939) tables apply to only one job; the problem of many jobs, as is the case for the military, will be addressed in the later chapters on cutscore analysis and simulating job assignments.

The Taylor-Russell (1939) tables can be used to interpret the validity coefficient as having "utility" in the context of an organization's cost associated parameters (e.g., current success rate, expected improvements given a more valid selection instrument, cost of maintaining the instrument, and difficulty in finding enough qualified candidates). However, Smith (1948) provided a cautionary note on using the Taylor-Russell tables. The assumption underlying the tables is bivariate normality no matter what the magnitude of the correlation. This theoretically requirement may not be met and there may be a mismatch in the Taylor Russell segment defined by the selection ratio and the empirical data at the selection ratio. That is, the selected segment of the bivariate normal distribution is only assumed to reflect the attributes of the operationally obtained empirical sample. In the military context, we cannot be sure that these two segments match, especially over time when the dynamics of the recruiting environment change where for some years there is an abundance of recruits with very high ASVAB scores because of a lack of private sector job opportunities or college costs, or conversely, a lack of high ASVAB scores when the economy improves.

Validity Coefficient Utility Interpretation: The Naylor-Shine Tables

The Taylor-Russell (1939) tables just discussed pertain to a dichotomous pass/fail (or successful/unsuccessful) criterion variable. The use of these tables assumes all personnel who pass the selection ratio cutscore perform at the same level. That is, there is no distinction in the relative contribution of employees who score highest on the predictor compared to those who score at or close to the cutscore. The Navy's philosophy is consistent with not making relative judgments about a Sailor's career performance potential at the training stage (where the tables are applied), because post training other personnel and organizational factors contribute over and above the ASVAB to job performance (e.g., OJT, motivation, career intentions, etc.). If, in the job context, the narrower view is to interpret the validity coefficient in differentiating personnel on their relative performance, the researcher could use the Naylor-Shine (1965) tables. The Naylor-Shine tables pertain to a continuous criterion variable and are used for a type of "expectancy" analysis where the interest is in determining and improving the mean predicted performance of employees. The Naylor-Shine tables are easy to use and require only the validity coefficient and a cutscore. The cutscore, however, must be identified in terms of a z-score within the sample, not projected for the population, so it is a sample-specific assessment. The basic equation of the Naylor-Shine tables is given by

$$\bar{z}_y = r_{xy} \frac{\varphi(z_x)}{SR}, \quad (3\text{-}1)$$

where $\varphi(z_x)$ is the ordinate of the normal distribution at the cutscore z_x, and SR is the selection ratio (e.g., Cascio, 1991).

As an example of the use of the Naylor-Shine (1965) tables, we assume two validity coefficients of .20 and .40, a base rate of .30, and a selection ratio (SR) of .40. A Taylor-Russell (1939) table analysis shows that a .20 validity coefficient is associated with a 37% success rate, whereas a .40 validity coefficient is associated with a 44 percent success rate. However, what does this 7% success rate improvement translate to in terms of improvement in mean performance? Remembering that the *z*-scores (standard score format) allow the validity coefficient to be the regression weight in this bivariate case, the predicted criterion would be twice as large applying the .40 validity than when applying the .20 validity. The computed expected mean criterion values (in terms of *z*-score) for the two correlation coefficients, .20 and .40, are .193 and .386, respectively.

In theory, the interpretation of the validity coefficient is that organizations can double the average performance of its employees merely by using a selection test that has twice as much predictive validity as the selection test in place. In practice, however, these performance gains are more difficult to achieve, because we are implicitly assuming that performance is strictly a function of the predictor. In the military training context, as in many other venues, training support systems are in place to deal with aptitude level declines (e.g., tutoring, night study sessions, and remediation). Also in the military job context, if job performance is the focus, performance is a function of several attributes, including performance in training, in which case a multi-attribute utility analysis may be more appropriate (Roth & Bobko, 1997). The main point that the Naylor-Shine equation makes is that one can expect increases in mean performance with larger validity coefficients. In the military training context, achieving twice as much predictive validity in the ASVAB world would not be possible given the already high ASVAB reliability and the academic/technical nature of the training criterion.

We refer the reader to Cascio (1993) for a full discussion of the role of utility in making selection decisions. Here, we describe briefly a well-known utility model that, as before, incorporates a selection instrument's predictive validity coefficient.

The Brogden-Cronbach-Gleser Utility Model

An expanded way to interpret the validity coefficient, and perhaps the most important to organizations, is in its role in estimating the cost savings, or *utility*, to the organization from implementing tests in their selection system. The Taylor-Russell (1939) and Naylor-Shine (1965) tables have been criticized in this regard. In particular, Cascio (1982) noted that a limitation common to both models is that "neither of these models formally integrates the concept of cost of selection or dollars gained or lost into the utility" (p. 222). The utility in the prior models discussed is strictly in terms of improved success rates and improved performance. However, the military is not a widget-producing organization, and so, it is difficult to tie military personnel productivity to cost savings due to a test's use in selection or classification. Nevertheless, we briefly discuss a model that does consider organizational dollar savings. Cronbach and Gleser (1965) built upon the work by Brogden (1959) to establish the Brogden-Cronbach-Gleser (BCG) model of utility. The BCG model shows how the validity coefficient of a selection instrument has practical consequences for an organization's productivity (utility in terms of dollar payoff, or ROI).

The BCG model formula is:

$$\Delta U = (N)(T)(SD_y)(r_{xy})(Z_x) - (N)(C_y)$$

where, with annotations referring to the military context,

ΔU = Increase in average dollar value payoff (in terms of employee productivity resulting from the valid selection process over random selection, where there would be no testing costs)

N = Number of job applicants to be tested (potentially a million military applicants per year)

C_y = Cost of testing one applicant (a factor for the military in a budget-constrained environment but not much more expensive than current costs to support the infrastructure required for applicant processing at the Military Entrance Processing Stations and the mental, moral, and physical checks)

T = Term of employment (varies from one term of enlistment to a full career)

Z_x = Average standard predictor score of the selected group at the ordinate of standard curve (height on the normal curve corresponding to the cutscore, which can be found in a statistical table or from a formula; this value is lower for more extreme cutscores)

SD_y = Standard deviation of dollar-valued job performance (projected for the normally distributed applicant population that was not subject to selection by a test with cutscore)

r_{xy} = Correlation between the selection test and performance measure.

As with the Taylor-Russell (1939) and Naylor-Shine (1965) tables, increasing the magnitude of the validity coefficient (r_{xy}), all other things equal, improves the utility of the testing program for the organization. Besides the military issues commented upon in the variable descriptions, another issue that has limited the BCG model's application in industry is how best to estimate the dollar standard deviation of job performance. We refer the reader to expanded applications of utility analysis (e.g., Cabrera & Raju, 2001; Johnson & Zeidner, 1991; Raju, Burke, & Normand, 1990).

Validity Coefficient Magnitudes Dependent on the Criterion

The DoD-sponsored Job Performance Measurement Project (JPM) discussed in the Introductory Manual involved various offshoot research projects. For example, the Center for Naval Analysis (CNA) conducted many studies on behalf of the Marine Corps. Carey (1992) conducted a study entitled "Does Choice of a Criterion Matter?" in which various surrogates of hands-on performance tests (HOPT; the National Academy of Sciences considered this type of performance measure to be the Gold Standard) were

tried out as alternatives because of the high cost of developing them across all military occupations. Carey showed that in the development of ASVAB classification composites used for classification to the Marine Corps infantry occupations, it did not matter much which surrogate criterion was used. (In fact, a criterion might not even be needed to develop the predictor composite; a rational approach that the Navy takes that involves linking the underlying ASVAB constructs to the training curriculum). However, if the ASVAB composite was to be used to develop selection standards, it definitely mattered which criterion was used to establish the ASVAB's validity. In a cautionary note, Carey stated that "…the choice of a criterion will make a difference when used for purposes where the strength of the relation between aptitude and criterion is critical. Setting selection standards is one such purpose" (p. 103).

Carey (1992) went on to show that the validity coefficient (corrected for range restriction) for the ASVAB General Technical (GT) composite (VE+AR) varied widely, depending upon the surrogate criterion that it was based upon. For example, the validity of the GT composite was .80 when the criterion was the occupation-tailored Job Knowledge Test; .42 for Grade Point Average in training, and .26 for the Supervisor rating (data from a Marine Corps Base labeled "B"). The various validity magnitudes would result in very different GT cutscores, some of which would not produce the 90% training success rate expected by the Marine Corps.

For example, in another paper, Carey and Wilbourn (1991) showed that by the "10 fail percent rule" a GT cutscore of 81 was required when the Job Knowledge Test was used as the predicted criterion versus a GT cutscore of 29 when the Supervisory Rating was used as the criterion. The broader conception of interpreting the validity coefficient amounts to deciding what performance context is most relevant for establishing selection standards. If the greatest concern for the military is training success (an expensive early career point of evaluation for the Navy), then measures of training success should be the criterion and the analysis of ASVAB validity coefficients is straightforward (final school grade becomes the continuous criterion variable).

If, on the other hand, the greatest concern is for predicting job performance, estimating the validity of the ASVAB, or any other predictor, cognitive or non-cognitive, becomes more complicated. An array of questions are then on the table such as (a) what is a sufficient magnitude for the validity coefficient given the it would most likely be much larger when training performance is the criterion, (b) what are the relative reliabilities of the measures (stability or equivalence), (c) how many recruits failed training and therefore are not in the job performance sample – a missing data problem, and (d) which job performance measures should you chose (e.g., supervisor ratings, job knowledge tests, or HOPT). Further, if predicting job performance is the goal, then given an already operational ASVAB standard that predicts training outcomes, one needs to consider how to structure the occupational classification model as a multiple hurdle that includes training performance (discussed in Chapters 15 and 16).

Obviously it is critical to understand the criterion and its relevance for establishing ASVAB standards. A recommendation coming out of one post JPM effort was to improve the final course grade to reflect what is learned and executed on the job (Sims & Hiatt, 2011). We can reflect on the conclusions stated by Mayberry and Carey (1993) from extensive work on the JPM project as follows:

> "Past validation research has typically concentrated on the identification of performance predictors (i.e., aptitude measures) with little regard for the quality, appropriateness, or completeness of the criterion measure. Such research usually collects the most convenient or readily available performance measures: supervisor ratings, training grades, promotion rates, etc. This is not to devalue such performance indicators, for each assesses some aspect of the multidimensional 'job performance construct…" (p. 39).

The Navy would endorse developing a final school grade that reflects not just knowledge learned but practical application of that knowledge.

Some Other Perspectives about Test Validity

Early in the industrial-organizational psychology measurement literature, Cronbach and Meehl (1955) and other testing experts recognized the issue of defining the term "validity." "Validity for what purpose?" was one of the questions addressed by the APA Committee on Psychological Tests charged with specifying "…what qualities should be investigated before a test is published" (p. 281). The multiyear project spelled out four types of validity that could all be interpreted as focusing on different criteria. Content validity focuses on establishing how well test items map to the specific areas intended to be measured without regard to an external performance criterion (e.g. math items in a predictor test map to math items that are also in the course assessment test). Construct validity focuses on how well the test captures the underlying domain of test performance that was hypothesized to relate to some external criterion (e.g., spatial visualization aptitude/ability hypothesized to relate to performance in an Air Traffic Controller tower. Criterion-related validity focuses on how well the predictor test actually relates to performance on a specific external criterion (and most likely assumes or knows that construct validity has been addressed in the test's development).

There are two criterion-related validity categories. The first is concurrent, where typically a new or experimental predictor test is tried out in a sample of individuals who have already met a cognitive test standard. These individuals typically are about to be measured on the performance criterion (if not at the same time as the predictor test's administration). Individuals may or may not be motivated when taking the new test as their performance may not be perceived as resulting in any decision about their status in the organization (as they are already hired). Therefore, the validity of the predictor in a concurrent validity setting may not reflect what would be found if the instrument were used operationally to make front-end hiring or job classification decisions.

The second criterion-related validity category is predictive, where the predictor test is administered at or nearly at the front end of the applicant assessment process. For the military, a new predictor test that has been evaluated positively in a concurrent validity setting may progress to a predictive validity setting (where the test could be administered alongside the CAT-ASVAB in the military screening process).

Cronbach and Meehl (1955) considered not only the type of validity considered but what it really means to establish the validity of a test: "One does not validate a test, but only a principle for making inferences. If a test yields many different types of inferences, some of them may be valid and others invalid…" (p. 297). The example presented earlier from the work of Carey (1992) and Carey and Wilbourn (1991) suggests that (a) calculation of the "right" predictive validity coefficient is important for the military and (b) criteria of convenience should be highly scrutinized. We refer the reader to Cronbach and Meehl for their classic report on construct validity and references to the APA committee's and other's published work and merely say that establishing the content and construct validity of the various ASVAB tests is the primary responsibility of the ASVAB test developer, Defense Manpower Data Center, Personnel Testing Division (DMDC-PTD); however, it is also the responsibility of the individual Services when they develop new tests intended for use in occupational classification as adjuncts to the ASVAB.

Concluding Remarks

This chapter discussed the interpretation of the validity coefficient assuming there are no factors that affect the calculation of its accuracy. The ASVAB validity coefficient of interest to the Navy in most cases applies to the ASVAB normative youth population from which future recruits are, theoretically, expected to be selected. Interpreting the ASVAB's validity coefficient (predictive, not concurrent when applied to the ASVAB) typically has involved demonstrating an expected improvement in the personnel success rates over random assignment from use of a predictor instrument. However, interpretation of the validity coefficient can also be expressed as (a) average expected improvements in performance scores and (b) proportional to the cost saving from replacing a selection/classification instrument with one that has improved validity (other factors held constant, such as cost of administration). Another way to interpret the validity of selection instruments is in terms of classification decision accuracy where, all other things being equal, larger magnitude of the validity coefficient will reduce classification errors. Finally, the criterion matters when establishing selection or classification standards and therefore establishing the appropriate criterion is part of the process of interpreting the ASVAB's validity coefficient. The next chapter provides a more in-depth discussion of the criterion and its reliability.

Chapter 3. References

Brogden, H. E. (1949). When testing pays off. *Personnel Psychology, 2,* 171-185.

Cabrera, E. F., & Raju, N. S. (2001). Utility analysis: Current trends and future directions. *International Journal of Selection and Classification, 9*, 92-102.

Carey, N. B. (1992). Does choice of a criterion matter? *Military Psychology, 4*, 103-117.

Carey, N. B., & Wilbourn, J. M. (1991). Setting standards and diagnosing training needs with surrogate job performance measures. *Military Psychology, 3*, 135-150.

Cascio, W. F. (1991). *Applied psychology in personnel management* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

Cascio, W. F. (1993). Assessing the utility of selection decisions: Theoretical and practical considerations. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 310-340). San Francisco, CA: Jossey- Bass.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana: University of Illinois Press.

Hull, C. L. (1928). *Aptitude testing*. London: Harrap.

Johnson, C. D., & Zeidner, J. (1991). *The economic benefits of predicting job performance Vol. II: Classification efficiency*. NY: Praeger.

Lawshe, C. H., & Boda, R. A. (1958). Expectancy charts I: Their use and empirical development. *Personnel Psychology*, *11*, 353-365.

Mayberry, P. W., & Carey, N. B. (1992). *Relationship between ASVAB and mechanical maintenance job performance* (CRM 92-192/March 1993). Alexandria, VA: Center for Naval Analyses.

Naylor, J. C., & Shine, L. C. (1965). A table for determining the increase in mean criterion score obtained by using a selection device. *Journal of Industrial Psychology*, *3*, 33-42.

Raju, N. S., Burke, M. J., & Normand, J. (1990). A new approach for utility analysis. *Journal of Applied Psychology, 75,* 3-12.

Roth, P. L., & Bobko, P. (1997). A research agenda for multi-attribute utility analysis in human resources management. *Human Resources Management Review,* 7, 341-368.

Sims, W., & Hiatt, C. M. (2011). *Promotion as a surrogate for job performance* (CAB D0024912.A1/Final, April 2011). Alexandria, VA: Center for Naval Analysis.

Smith, M. (1948). Cautions concerning the use of the Taylor-Russell tables in employee selection. *Journal of Applied Psychology, 32,* 595-600.

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology, 23*, 565-578.

Chapter 4.
Measurement Error and Reliability Estimators
Sarah A. Hezlett

Introduction

Because it is not possible to develop psychological measures to be perfectly reliable, the variables of interested in our studies are imperfect reflections of their theoretical or "true" relations. We saw in Chapter 2 that measurement errors can affect correlations, as can a host of other reasons, including restriction in range of predictor test scores due to operational use of a cutscore. In this chapter, we expand on the measurement error topic keeping in mind that while measurement error is random at the individual level, it systematically affects the accuracy of the estimation of the validity coefficient. Further, some types of error are systematic and contribute to the true score variance while other types – the kind we attempt to measure, are random errors. This chapter also provides a few examples of reliabilities that apply to different types of criterion and some that are reported for the ASVAB tests. Finally, we cite some critiques in the literature of studies that have misinterpreted the measurement error situation (Schmidt & Hunter, 1996).

Background

As discussed in Chapter 2, measurement error reduces the reliability of any psychological measure and therefore places a limit on the magnitude of the correlation between a predictor and criterion measure. However, in general, validity coefficients should be corrected for measurement error only when researchers are attempting to understand how the constructs measured by two or more measures are related (Society for Industrial and Organizational Psychology, 2003), which would include any experimental predictor considered for addition to the ASVAB and its relation to the criterion. In the operational setting, military applicants' ASVAB scores of record, and not their "true" scores (indeterminate) are the scores used to make enlistment and occupational classification decisions, therefore we do not correct the ASVAB (the predictor) for unreliability.

For the criterion variable, test reliability is of concern in both the research and operational setting. In the research setting, Mayberry and Wright (1992) paid particular attention to criterion reliability during the joint-service Job Performance Measurement (JPM) project (the 1980s project discussed in the Introductory Manual). Mayberry and Wright saw three reasons why we would want to have reliable measures: (a) the need to have consistent and meaningful measurements for an individual, (b) to be able to generalize a single measurement to a larger context; and (c) to avoid limiting the validity of the measure used to predict the criterion. An additional reason for wanting a reliable criterion measure when conducting operational ASVAB validation/standards studies is that if found to be unreliable, there may be a need to recommend a complete rework of the measure to ensure a, b, and c above.

For the most part, the Navy's training command (Navy Education and Training Command, NETC) has high standards in developing the criterion used in ASVAB validity/standards, the school performance tests in technical training. However, if the ASVAB validity coefficient resulting from use of the criterion is much lower than expected from historical studies, then further exploration of the criterion would be recommended in a follow-on study (as well as addressing any ASVAB compromise that may have occurred). The high validity once known for a predictor in this case has been affected and there is question about the instrument's usability for future selection/classification decisions.

Navy research has not been conducted in recent years to assess the reliability of the "final school grade" variable used in in Navy training and so we cannot state a range of plausible values. We note, however, that Hunter (1986) and Hunter and Hunter (1984) apparently conducted reliability research on Navy training performance criteria (also reported in Salgado, Anderson, Moscosco, Bertua, & De Fruyt, 2003) that tend to be objective tests. The estimated average criterion reliability reported from this work (meta-analysis) is about .80. The method used to estimate the criterion reliabilities is not known (to the current authors) but the .80 value is consistent with what Mayberry and Wright (1992) found for objective job knowledge tests administered to two Marine Corps mechanical occupations (an average of about .80) where the method of estimating reliability was test retest. We note that .90 reliability coefficient appears to be sort of a "gold standard" (Guilford & Fruchter, 1978; Nunnally, 1978), but that reliabilities of lower magnitude can still be adequate for many selection and classification situations.

Experimental predictors being evaluated in a research context should be submitted to the process of correcting their validity coefficients for errors in measurement, both in the predictor and the criterion. That is, in the "theoretical" world, as mentioned, the interest may be solely in establishing the worthiness of the predictor instrument. If a criterion variable has low reliability, it will obscure the predictor's value (remembering reliability puts a ceiling on validity). Or, the experimental predictor itself may have less than satisfactory reliability because, for example, there were limited resources applied in the test development process. Documenting the issues that might have led to low predictor validity and also providing a correction for reliability that shows the potential for the instrument will possibly lead to issue resolution, especially if past research supports the underlying construct purported to be measured by the experimental predictor measure is linked to jobs (by job analysis). (Chapters 6 and 7 deal with joint corrections for reliability and range restriction.)

It is not possible to know for sure what kinds of influences cause a predictor or criterion measure's scores to be unreliable, or to know an individual's "true" score. The general broad factors that contribute to the unreliability of a measure's scores are the testing environment (e.g., not standardized), the examinee's state of mind (e.g., unmotivated or tired), the appropriateness of the examinee population (e.g., the test may be too difficult or easy for the target population), and the test itself (adequate resources and expertise might not have been available to develop a psychometrically acceptable test).

It is also not possible to know the exact amount of error in any psychological measure and so error must be estimated (see formulas presented in Chapter 2 and later in this chapter). A variety of indices of measurement error have been developed, each with its own strengths and assumptions. The kinds of reliability coefficients discussed in this chapter are not exhaustive but address the most pertinent that the ASVAB validation/standards researcher will want to consider: (a) internal consistency (do the items correlate highly and similarly with one another, indicating that scores on each item provide you with similar information about the examinee), (b) stability (does the test produce the same correlation for examinees tested twice over some meaningful time between testing sessions, assuming no learning has occurred), (c) equivalence (to what degree do parallel forms of a test correlate), (d) interrater (do the same panel members rate individuals the same if this is a method of performance evaluation), and (e) intrarater (does the same rater grade all individuals according to the structured grading methods and not let influencers such as fatigue over time or perceptions of a student that biases the grading process).

Defining Reliability

As was shown in Chapter 2, the concept of measurement error is captured in a simple equation that is a central tenet of classical test theory (CTT). In this section, we revisit the development of CTT reliability equations with sometimes a different perspective than in Chapter 2.

The CTT reliability equation expresses the value of the variable that is observed (x) as a function of the true value of the variable (T) and measurement error (e).

$$x = T + e \quad (4\text{-}1)$$

CTT, sometimes referred to as true-score theory, is not the only theoretical approach to understanding psychological measurement, but it is one of the oldest and most dominant approaches. Although some of CTT's implications regarding the selection of reliability estimates diverge from those of other psychometric theories, for example, domain sampling (where any measure if considered a compilation of a random set of items from a specific content domain – Nunnally's (1998) favored model or parallel tests, the conclusions about the nature of tests and true scores are all the same (Ghiselli, Campbell, & Zedeck, 1981). We reiterate the point (from Chapter 2) that all psychometric theories differentiate systematic biasing errors such as "test-wiseness" and random errors such as examinees' moods at testing time. Random errors are assumed uncorrelated to true scores and are treated as an additive component to the true score component in our observed score. Systematic error, on the other hand, is considered mathematically as part of the true score and can therefore increase reliability because there is an increase in the proportion of true score variance to total test score variance. Systematic bias, however, can affect test validity**.** Although there are methods for disentangling the systematic error from the unsystematic random error, they are not discussed in this manual.

It is also important to know that when we refer to an instrument's reliability, we are not referring exactly to that instrument itself, but the measurements that are made by that instrument, which may be population and sample specific. In the ASVAB High School Testing Program (STP) (now called the Career Exploration Program – CEP), norms and test reliabilities are reported by grade and gender.

Reliability estimates, as we know by now, are used to quantify the amount of measurement error in an observed variable or measure (Charles, 2005). Conceptually, reliability refers to the extent to which an individual scores the same way when measured multiple times (Ghiselli et al., 1981). Perfect reliability (r_{xx} = 1.0) would occur only if an individual always received the same score on a measure (or equivalent set of measures) no matter when it was (they were) administered (assuming no learning has occurred). Under CTT, multiple tests are parallel when they have (a) equal means, (b) equal standard deviations, (c) equal correlations with each other, and (d) equal correlations with scores on any other measure (Ghiselli et al.).

The development of the concept of reliability within CTT that includes parallel tests (of which many have been developed in the ASVAB testing program) is grounded on three assumptions (Ghiselli et al., 1981). The first assumption is captured in Equation 4-1: Observed scores are an additive function of true and error scores. The second assumption clarifies the nature of true scores in that it assumes that individuals have stable characteristics that persist over time. The third assumption clarifies the nature of measurement error in that it assumes that error is completely random and, therefore, independent of and uncorrelated with all other characteristics.

Ghiselli et al. (1981) showed that these three assumptions and the concept of parallel tests can be combined to derive a mathematical definition of reliability. Assuming variables are in deviation score form, the correlation between two parallel tests x_1 and x_2 is

$$r_{x_1x_2} = \frac{\sum x_1 x_2}{n\sigma_{x_1}\sigma_{x_2}}. \quad (4\text{-}2)$$

Because an individual's true scores on a series of parallel tests are equal and because the standard deviations of all parallel tests are equal, substituting Equation 4-1 into Equation 4-2 and replacing $\sigma_{x_1}\sigma_{x_2}$ with σ_x^2 results in

$$r_{x_1x_2} = \frac{\sum (T + e_1)(T + e_2)}{n\sigma_x^2}. \quad (4\text{-}3)$$

This expression can be expanded to,

$$r_{x_1x_2} = \frac{\sum\left(T^2 + Te_1 + Te_2 + e_1e_2\right)}{n\sigma_x^2} \quad (4\text{-}4)$$

and re-arranged to,

$$r_{x_1x_2} = \frac{\frac{\Sigma T^2}{n} + \frac{\Sigma Te_1}{n} + \frac{\Sigma Te_2}{n} + \frac{\Sigma e_1e_2}{n}}{\sigma_x^2}. \qquad (4\text{-}5)$$

Because the correlation between two variables multiplied by their standard deviations is equal to the sum of the cross products of deviation, this equation can be re-written as follows:

$$r_{x_1x_2} = \frac{\sigma_T^2 + r_{Te_1}\sigma_T\sigma_{e_1} + r_{Te_2}\sigma_T\sigma_{e_2} + r_{e_1e_2}\sigma_{e_1}\sigma_{e_2}}{\sigma_x^2}. \quad (4\text{-}6)$$

By the assumption that errors are random, the correlation between error terms and other variables is zero. Therefore, Equation 4-6 reduces to

$$r_{x_1x_2} = \frac{\sigma_T^2}{\sigma_x^2}. \qquad (4\text{-}7)$$

Thus, to reiterate, the reliability is the proportion of the variance observed in a measure that is attributable to true variations in the value of the variable (Charles, 2005; Ghiselli et al., 1981). Traditionally, r_{xx} has been used to denote the reliability of predictors, while r_{yy} has been used to indicate the reliability of criteria. Estimates of reliability typically range from 0.0 to 1.0, with values close to 1.0 indicating greater reliability and therefore measurement with less error.

As an example of the effect of the magnitude of an estimated reliability coefficient on the size of the corrected validity coefficient, suppose the observed correlation between a predictor and a criterion is .30 but we assume that it is attenuated as a result of criterion unreliability. If the estimated reliability of the criterion is .52, the corrected validity coefficient will increase to .42. On the other hand, if the reliability is estimated to be a much greater .86, the corrected validity coefficient will increase to only .32. The closer the reliability estimate is to 1.0, the less of an improvement in the validity coefficient resulting from the correction for unreliability. We caution that we would want to follow the advice of experts, including Nunnally and Bernstein (1994), that a reliability coefficient of .70 may be considered sufficient for internal consistency measures in the initial development of a measure, but that in operational use .80 or even .90 may be required.

Selection of Reliability Estimates

Historically, there has been debate over what type of reliability estimates should be used in making corrections to the validity coefficient for attenuation due to unreliability (Muchinsky, 1996). The general convention is to use a reliability estimate that is aligned with one's "views" of what are the most important sources of error in a particular validation situation (Muchinsky). To have a "view", it is important to have an understanding of common kinds of reliability estimates and the sources of measurement error they capture.

Common indices of reliability include estimates of internal consistency reliability (e.g., Cronbach's Alpha, KR-20), test-retest reliability (i.e., stability), the coefficient of equivalence and stability, interrater reliability, intrarater reliability, and coding reliability. Before discussing the sources of error captured in each of these kinds of reliability estimates, it is important to note that it is not appropriate to make repeated corrections of a validity coefficient using multiple reliability estimates. That is, the validity coefficient should not be first corrected using an internal consistency estimate of reliability, then a stability coefficient, and then an estimate of interrater reliability.

The practitioner should select the single most appropriate reliability estimate available and perform one validity coefficient correction for attenuation (correction for unreliability). We refer the reader to Nunnally's (1978) classic text on CTT for, among other important topics, the chapter on "Theory of Measurement Error." The description of the types of reliability described below are from the perspective of Hunter and Schmidt (1977; 2004) who were concerned about the various types of random effects that would affect a meta-analysis procedure (and the generalization of an identified effect across samples although not within the realm of an ASVAB validation/standards study). For those familiar with the framework of validity generalization (VG) that considers multiple studies and multiple statistical artifacts, there is only one type of measurement considered, purely random error, not the systematic error introduced by the factors stated in the previous paragraph.

Internal Consistency Reliability

The terms internal consistency reliability traditionally has been used to describe reliability estimates that are based upon individuals' responses to multiple, independent stimuli during a single measurement session. The multiple stimuli are assumed to be parallel, making them equivalent in assessing the variable of interest (Hunter & Schmidt, 2004). Examples of criterion measures that involve multiple responses during a single measurement session include multiple-item tests of job knowledge and job simulations or samples that require incumbents to react to several independent events.

Having multiple responses completed at a single point in time permits us to identify two types of measurement error: (a) random response error and (b) specific error (previously referred to in this chapter as systematic error) (Hunter & Schmidt, 2004). Random response error has been characterized as "noise" in the human nervous system (Hunter & Schmidt). It involves arbitrary behaviors not systematically related to any characteristics of the person being assessed or the situation in which the person is being

measured. Specific response errors are errors stemming from reactions to a particular situation or stimulus (e.g., item content). Technically, these errors are person-by-item interactions (see Hunter & Schmidt, p. 100).

Internal consistency reliability coefficients fail to capture a third potential source of measurement error (Ghiselli et al., 1981; Hunter & Schmidt, 2004) -- transient error. Transient errors are due to factors that vary randomly over time, such as mood or illness. Because internal consistency reliability estimates are based on the collection of data at a single point in time, these coefficients do not detect transient errors (Ghiselli et al., Hunter & Schmidt). Therefore, internal consistency reliability coefficients may overestimate the reliability of a criterion measure if transient errors are present. Consequently, using internal consistency reliability coefficients, such as Cronbach's alpha (Cronbach, 1951) or KR-20 formula (Kuder-Richardson, 1937), to correct for attenuation may lead to underestimates of the operational validity of a set of predictors.

Test-Retest (Stability) Reliability

Test-retest reliability estimates are computed by correlating individuals' responses to a stimulus (e.g., a job knowledge test, a job simulation) at two points in time (Ghiselli et al., 1981; Hunter & Schmidt, 2004). For example, a group of job incumbents might complete a job simulation soon after they complete their job training (Time 1) and then complete the same job simulation several weeks later (Time 2). The correlation between the incumbents' scores at Time 1 and Time 2 provides an estimate of the simulation's stability measure of reliability.

For test-retest reliability estimates, one needs to be careful not to take the two measurements very far apart in time as learning might occur for some but not all subjects, thus shifting the rank ordering of performance from the original order. This shift in rank ordering at Time 2 from Time 1 thus lowers the magnitude of the calculated reliability coefficient (merely, the correlation between Time 2 and Time 1 scores).

Test-retest reliability takes into account two sources of measurement error: (a) random response error, and (b) transient error. The correlation of individuals' scores at two different times will be decreased to the extent that individuals' responses at either time are affected by random "noise" or by transitory factors, such as mood, fatigue, or illness (Ghiselli et al., 1981). On the other hand, specific error is not assessed by estimates of test-retest reliability. Because the same stimulus is administered at both time periods, the same measurement errors that arise from specific aspects of the stimuli (e.g., specific instructions, test items) will occur in both time periods. The stability coefficient, therefore, will be too large because this common specific error will appear to be true score variance (Hunter & Schmidt, 2004). As with differential learning over time, practice effects might affect the reliability estimate (Ghiselli et al.) if they are not uniformly applicable to all applicants (e.g., differential practice effects as with differential learning will lead to a lower test-retest reliability estimate).

Coefficient of Equivalence

The coefficient of equivalence considers all three kinds of measurement error (stability, internal consistency, and equivalence) but is largely concerned with the latter. Computing this reliability estimate requires administering two parallel forms of a stimulus at two points in time (Hunter & Schmidt, 2004). This is the most difficult type of reliability to assess because of the difficulty in creating strictly parallel test forms that are equivalent in means, standard deviations, and other moments of their score distributions (Ghiselli et al., 1981). For paper and pencil ASVAB forms, parallel form development efforts have occurred over years that involve item writing, tryout of these items in a nearly full range population of youth, and form assembly based upon item parameters that are matched across forms. The computer adaptive version of the ASVAB, CAT-ASVAB, develops pools of items based upon Item Response Theory and parameters with item selection for an individual based upon an adaptive algorithm.

Interrater Reliability

In many validation studies, the criterion is developed specifically for the study and measured by having observers evaluate the individuals constituting the validation sample. For example, supervisors or peers may rate job incumbents' performance or evaluators may rate trainees' proficiency. The sources of error in such judgments can be clustered into two broad categories: (a) error in judgment and (b) idiosyncratic rater perceptions (Hunter & Schmidt, 2004). Because judgments are responses to stimuli, they (like responses to items on a test) are affected by random error, specific error, and transient error (Hunter & Schmidt). For example, a supervisor's ratings may be affected by random "noise," by idiosyncratic errors induced by the wording of the rating scales, or by errors in judgment generated by temporary influence such as the supervisor's illness, mood, or fatigue. In addition, raters have their own idiosyncratic biases that affect perceptions of others and these biases are thus not part of the criterion construct and, consequently, are a form of error (Hunter & Schmidt).

The correlation between ratings of the same people provided by different raters serves as an estimate of interrater reliability. This estimate of reliability takes into account both errors of judgment and idiosyncratic rater perceptions (Hunter & Schmidt, 2004). The correlation between the judges will be reduced to the extent that either judge responds randomly, makes errors related to the specific scales being used, independently experiences a source of transient error, or has idiosyncratic perceptions (Hunter & Schmidt). Thus, the inter-rater reliability coefficient is analogous to the coefficient of equivalence and stability (Hunter & Schmidt). It is generally agreed that interrater reliability is the appropriate way to estimate reliability for judgments; however, we will see later that intrarater reliability is generally higher.[2]

[2] Note that interrater reliability (which considers the similarity of the rank ordering of ratees by raters) differs from interrater agreement (which considers the similarity of the absolute magnitude of the ratings provided to ratees across raters) (Tinsley & Weiss, 1975).

Intrarater Reliability

At times, data to compute interrater reliabilities are not available. For example, it might not be feasible for individuals in the validation sample to be evaluated by multiple raters. When there is only a single set of ratings for each individual, an alternate reliability estimate can be computed if individuals have been rated on more than one dimension (e.g., multiple dimensions of performance, or multiple items assessing training success). This intrarater reliability is analogous to an internal consistency reliability estimate (Hunter & Schmidt, 2004). It captures only random errors and specific errors; it does not assess either transient error or idiosyncrasies in rater perceptions (Hunter & Schmidt, 2004). Not surprisingly, estimates of intrarater reliability are often much higher than estimates of interrater reliability (Visweswaran et al., 1996).

Coding Reliability

Occasionally, coded data are used as a criterion measure. For example, trainees' performance on a leadership simulation is recorded. Later, the trainees' non-verbal behavior is scored by trained observers. Coding discrepancies between the trained observers are known as coding error (Hunter & Schmidt, 2004). There are multiple ways of estimating the reliability of coding, including computing the correlation between different trained observers' coding.

Coding error is important to assess, but it is not the only source of measurement error that affects coded data (Hunter & Schmidt, 2004). It also is critical to consider the sources of error in the behavior being coded. For example, a trainees' performance on a leadership simulation is likely to be affected by random error, specific error, and transient error. These sources of error are not captured by the extent to which trained observers' codings agree, making coding reliability a potentially poor choice for corrections for attenuation (Hunter & Schmidt, 2004). In some cases, coding error might be very low; resulting in high estimates of coding reliability, but the consistency of the behavior being sampled might be quite low. In these instances, coding reliability would substantially overestimate the actual reliability of the criterion of interest and under-correct the validity coefficient, resulting in a downwardly biased estimate of the predictor's validity.

Meta-analytic Sources of Job Performance (Criterion) Reliability Estimates

In using meta-analytic estimates of reliability in correcting validity coefficients for measurement error, it is critical to select reliability estimates that are well-aligned with the criteria used in the current validation study. For example, if the criterion of interest is supervisory ratings for overall job performance, then the reliability coefficient should be one developed on the same performance metric. Meta-analytic studies will produce an "average" reliability with a range and standard deviation. Taking the average as the reliability coefficient that "may" apply to one's study could be the safest thing for a researcher to do remembering that there should be similarity in the content of the criteria, the length of the measurement instrument, and the number of raters.

Table 4-1 gives examples of meta-analytic reliability estimates (averages) where job performance served as the criterion in all cases but with three ways of estimating the reliability coefficient (Interrater, Intrarater, and Stability).

Table 4-1
Meta-Analytic Estimates of Job Performance Measures' Reliability

Criteria	Type of reliability	*k*	*N*	r_{yy}	SDr_{yy}
Job perf., supervisor ratings[a]	Interrater	40	14,650	.52	.10
Job perf., supervisor ratings[a]	Intrarater	89	17,899	.86	.14
Job perf., supervisor ratings[a]	Stability	12	1,374	.81	.09
Job perf., peer ratings[a]	Interrater	9	2,389	.42	.11
Job perf., peer ratings[a]	Intrarater	10	1,270	.85	.12

Note. K = the number of studies. Reliabilities are for incumbents.
[a]Visweswaran et al. (1996) also report reliabilities for different aspects of jobs performance such as leadership, job knowledge, and effort.

Table 4-1 shows that the Interrater reliabilities for both supervisory and peer ratings (.52 and .42, respectively) are much lower than the Intrarater reliabilities for the same (.86 and .85, respectively) (as noted they should be in a previous sub-section on Intrarater Reliability). Second, the magnitudes of the standard deviations of reliability estimates across studies for every type of reliability highlight the variability in coefficients of reliability. In this regard, it has been argued that the highly cited and used mean interrater reliability for supervisory job performance ratings estimated by Visweswaran et al. (1996) might be a lower bound estimate of reliability (Scullen, Bergey, & Aiman-Smith, 2005). We note that wide use of the value indicates that in the published work on the reliability of the supervisory ratings, the Interrater reliability is deemed most appropriate.

Marine Corps Job Performance (Criterion) Reliability Estimates

As mentioned earlier in this chapter and at various points in both the Introductory and Technical Manuals, the joint-service Job Performance Measurement (JPM) project was a huge endeavor to ascertain the magnitude of the predictive relationship between the ASVAB and the job performance. For the most part, the project bypassed ASVAB's relationship with training performance and concentrated heavily on the development of job performance measures of various types focusing on the gold standard, hands-on measures of job performance (Wigdor & Green, 1991). Each Service was responsible for developing the performance criteria for a number of occupations that were selected to be reflective of a divers set of occupational areas. In retrospect, it appears that the Marine Corps JPM efforts were the most in-depth in the development and publishing of their work.

Table 4-2 has been taken from Mayberry and Wright (1992) (Table 1 and Appendix Table A-1) and contains reliability estimates for hands-on measures of job performance and job knowledge measures that apply to five helicopter mechanic military occupations (MOS) (different helicopter platforms).

Table 4-2
Sample Derived Criterion Reliability Estimates from the Marine Corps JPM Project (Mayberry & Wright, 1992)

	Military Occupational Specialties (1980 Reference Form 8a)				
	Job-1	**Job-2**	**Job-3**	**Job-4**	**Job-5**
Hands-on performance test					
Test-retest	.70 (.79)	.81 (88)	-	-	-
Split-halves	.71 (.80)	.80 (.87)	.85 (.91)	.74 (.82)	.84 (.87)
Alpha coefficient	.73 (.81)	.81 (.88)	.81 (88)	.69 (.78)	.77 (.81)
Job Knowledge test					
Test-retest	.61 (73)	.77 (.87)	-	-	-
Split-halves	.91 (.94)	.95 (.97)	.93 (.96)	.92 (.95)	.91 (.92)
Alpha coefficient	.90 (.93)	.95 (.97)	.92 (.96)	.90 (.94)	.92 (.93)

Notes. (1) Jobs are helicopter mechanics for four different platforms. (2) Range corrected reliabilities are shown in parentheses. (3) Score agreement reliabilities documented in the report are not included in the table.

Mayberry and Wright (1992) note that the first reliabilities of the pair listed in Table 4-2 were developed in range sample available for the four helicopter mechanic MOSs. These reliabilities are restricted in range due to the operational use of an ASVAB standard (composite with cutscore) in the sense that the full range of ASVAB scoring youth were not considered in the measures' development – just those who met the ASVAB classification standard, graduated from training, and reported to the job. The range corrected reliabilities are listed in parenthesis and we defer discussion of the range correction process until Chapters 6 and 7.

Although the military most likely will never see the level of effort that went into criterion development during the JPM days, we caution about accepting criterion of convenience that most researchers have access to in their organization's databases without questioning how they were derived. (Chapter 4 of the Introductory Manual addresses the criterion problem and the potential pitfalls of criteria of convenience.)

Paper and Pencil and CAT-ASVAB Reported Reliabilities

The developers of the ASVAB, Defense Manpower Data Center, Personnel Testing Division (DMDC-PTD) report estimated reliabilities of the ASVAB tests from Item Response Theory (IRT) methods, not CTT methods. The IRT method provides an analogue to the CTT methods, called marginal reliability, and involves averaging expected error variance across the ability (theta) range and transforming them in to reliability coefficients (conceptually, internal consistency reliabilities). DMDC provides a link to the Official ASVAB website that includes many relevant research documents, one of which is an explanation of the IRT methods, references, and reported CAT-ASVAB reliabilities www.officialasvab.com/reliability_res.htm . We site only one of DMDC's references related to reliability estimation using IRT methods (Sireci, Thissen, & Wainer, 1991). Reliabilities for paper and pencil ASVAB forms are also published (e.g., Palmer et al., 1988; Sands, Waters, & McBride 1997). Table 4-3 is provided to give the reader an idea of the magnitude of the ASVAB test reliabilities for both test-retest (stability) and parallel forms (equivalency).

Table 4-3
Test Retest and Parallel Forms Reliabilities for Earlier ASVAB Power Tests

	ASVAB Power Tests (1980 8a Reference Form)							
	GS	**AR**	**WK**	**PC**	**AS**	**MK**	**MC**	**EI**
Test-Retest	.83	.91	.91	.78	.86	.89	.83	.79
Parallel Forms	.82	.88	.90	.79	.82	.87	.78	.75

Notes. (1) See Chapter 2 of the Introductory Manual for descriptions of the ASVAB tests. (2) Reliabilities were developed on 12th graders participating in the High School Student Testing Program – now called the Career Exploration Program (CEP). (3) All values were taken from the ASVAB Technical Manual for the ASVAB 18/19 CEP, which are published elsewhere.

Table 4-3 shows some similarities between test-retest and parallel forms reliabilities for the individual subtests. Larger differences are noted between tests (e.g., PC and WK – both part of the ASVAB Verbal composite). When WK and PC are formed into a composite (Verbal [VE]), the verbal construct is more reliably measured.

Measurement Error Scenario Observations

Schmidt and Hunter (1996) were concerned that many researchers who publish validity studies have misconceptions about measurement error and how it should be addressed. The authors provided a critique of 26 studies under different scenarios with the intent of augmenting the mostly theory and formula-based treatments of the measurement error topic. Of the 26 scenarios, we discuss only two and note that all are worth reading.

Scenario 6 (Measurement error not addressed in theory-based aptitude research)

Schmidt and Hunter (1996) described Scenario 6 for a researcher who tested specific aptitude theory against the *g* theory using three tests that measured quantitative, verbal, and spatial ability, all three having reliabilities of at least .80 (not specified if test-retest, parallel forms, or internal consistency). The criterion was a measure of job performance (not specified what aspect of job performance, overall, or if reliability was measured and how). In the regression analysis, *g* was entered first into the equation to predict job performance, followed by the three individual tests. It appears that two of the three tests had standardized regression weights that were statistically significant and of practical value. However, Schmidt and Hunter pointed out that Schmidt, Hunter, and Caplan (1981) have discussed this phenomenon in detail and that what is missing from the researcher's method was to correct the criterion related validity for measurement error (we presume just the predictors could be corrected). When the corrections were subsequently done, the standardized regression weights that were originally significant became zero, supporting the g theory.

Specific aptitude theory versus general mental ability in psychological research has been an important topic in the ASVAB community for a long time (e.g., Ree & Earles, 1991; Ree, Earles, & Teachout, 1994). That is, should we just be using the general factor "*g*" taken as the first principle component factor as the measure of the full ASVAB's utility in predicting training or job performance or should we use occupation tailored ASVAB composite comprised of several tests (instead of all ASVAB tests in measuring *g*) to emphasizing differential assignment and person/job fit. Using all tests in a measure of *g* boosts the reliability, and therefore, potentially boosts validity across many occupations whereas using tailored composites would have lower validity.

Scenario 10 (Validity correction using the wrong reliability coefficient for supervisory ratings)

Schmidt and Hunter (1996) described Scenario 10 for a researcher who was interested in the predictive validity of a personnel assessment outcome measure when the criterion was the supervisor's ratings of job performance. There were 10 job dimensions for a single supervisor to rate and an overall "index" of job performance was the sum of standardized scores across all dimensions. Coefficient alpha, derived from the intercorrelations of the 10 dimensions, was used to correct the validity coefficient for unreliability in the criterion measure. Schmidt and Hunter point out that coefficient alpha (internal consistency) is inappropriate because, basically, with one rater, the idiosyncrasies of the rater and his/her perceptions of the rated individual would somewhat affect the ratings on all dimensions. Mathematically, this systematic error becomes part of the true score variance and increases the ratio of true to observed score variance with the result of an upwardly biased validity coefficient. Schmidt and Hunter state about the intrarater reliability that "It is an estimate of what the correlation would be if the same rater rerated the same employees …"p. 209. The authors go on to cite Cronbach (1951) on this fundamental issue.

Schmidt and Hunter (1996) state that the appropriate reliability in this scenario is interrater reliability (which would require a rework of the research design) and cite others who have documented the large specific rater error (between raters) particularly when the criterion is job performance (King, Hunter, & Schmidt, 1980; Rothstein, 1990). The finding of low intrarater reliability and high intrarater reliability in these researchers' work is consistent with what we observed in Table 4-1 of this chapter. We have included this scenario because supervisory ratings are a part of both military and private industry performance evaluation systems.

Scenario 15 (Construct equivalence of speeded tests using wrong reliability coefficient)

Schmidt and Hunter (1996) described Scenario 15 for a researcher who was directed by his organization to develop an in-house clerical speed/accuracy test that was construct equivalent to a commercially published test (Minnesota Clerical Test). The in-house developed test was administered alongside the commercial test for over 1,800 job applicants – the scores from the two tests correlated .81. KR-20 (internal consistency) estimates of reliability were .96 for the in-house measure and .94 for the commercial. Dividing .81 by the produce of the square roots of .96 and .94 yielded .85. This estimate of the true score correlation was not of sufficient magnitude to conclude that the two measures were construct equivalent.

Schmidt and Hunter (1996) pointed out that KR-20, a special case of coefficient alpha (internal consistency), was not appropriate for use with speeded tests because all of the items in speeded tests are typically of the same type (homogeneous). Further, the items are so easy that if given enough time, all examinees could answer all of the items correctly (thus inflating reliability). Schmidt and Hunter suggested a parallel forms approach to estimate construct equivalence and the researcher followed up by (a) splitting each test in half and administering two timed halves for each and (b) correcting the two resulting intercorrelations with the Spearman-Brown (Brown, 1910; Spearman, 1920) formula, thereby estimating reliabilities for length corrected forms (doubling each test to their original length). The resulting reliabilities were .79 and .87 (new and commercial instruments), which yielded an estimated .98 correlation between the true scores of each measure – confirming construct equivalency.

As an aside, we note that the former ASVAB clerical speed/accuracy test, Coding Speed, has been found to add incremental validity to the ASVAB for not only clerical types of jobs, but for other types of Navy Ratings such as Air Traffic Controller and the Navy SEALs (Appendix A of the Introductory Manual). From time to time, the question comes up within the ASVAB community about what is actually being measured by Coding Speed. Segal (2012) through extensive analyses concludes that there is a motivational component.

Concluding Remarks

This chapter provided a discussion of the underlying tenants of test measurement error and the various types of reliability estimates that apply to both the predictor and the criterion. In our focus on the criterion, we included reliability averages taken from the meta-analytic literature and showed the differences in the magnitudes of reliability coefficients for three methods of measuring job performance (interrater, intrarater, and stability). We discussed the ASVAB reliabilities for both CAT-ASVAB (IRT methods) and paper and pencil ASVAB. We again note that the ASVAB selection and occupational classification composites containing more than one individual test result in higher reliability coefficients, about .90 or above. However, we do not correct for predictor reliability in operationally focused ASVAB validation/standards studies. Currently, the not even ASVAB composite validities are corrected for measurement error in the criterion variable, which typically is the training grade variable, also because of the operational focus. However, the position of ignoring reliability in the criterion may change as the military experiences more protracted financial constraints placing at risk the reliability of all types of military performance measures. Further, as personality and interest measures are considered for selection and classification decisions, the underlying true relationships of predictors and post training performance measures (where personality theoretically applies) will require consideration of the appropriate reliability indices and correction methods; even appropriate indices/methods should be questioned for use in the operational context.

Chapter 5 goes in depth on the topic of range restriction and the various corrections of the validity coefficient derived in a selected sample. Chapters 6 and 7 address the joint correction for restriction in range and measurement error (reliability).

Chapter 4. References

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3,* 296–322.

Charles, E. P. (2005). The correction for attenuation due to measurement error: Clarifying concepts and creating confidence sets. *Psychological Methods, 10,* 206-226.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16, 297-334.*

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences.* San Francisco: W. H. Freeman and Company.

Guilford, J. P., & Fruchter, B. (1978). *Fundamental statistics in psychology and education.* NY: McGraw Hill.

Hunter, J. E. (1986). Cognitive ability, job knowledge, and job performance. *Journal of Vocational Behavior, 29,* 340-362.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96,* 72-98.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis* (2nd ed.). Thousand Oaks, CA: Sage.

King, L. M., Hunter, J. E., & Schmidt, F .L. (1980). Halo in a multidimensional forced-choice performance evaluation scale. *Journal of Applied Psychology, 33, 507-516.*

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2,* 151–160.

Mayberry, P. W., & Wright, W. H. (1992). *Reliability of mechanical maintenance performance measures* (CRM 91-246/February 1992). Alexandria, VA: Center for Naval Analyses.

Muchinsky, P. M. (1996). The correction for attenuation. *Educational and Psychological Measurement, 56,* 63-75.

Nunnally, J. (1978). *Psychometric theory* (2nd ed.). NY: McGraw-Hill.

Palmer, P., Hartke, D. D., Ree, M. J., Welsch, J. R., & Valentine, L. D., Jr. (1988). *Armed Services Vocational Aptitude Battery (ASVAB): Alternative forms reliability (forms 8, 9, 10 and 11)* (AFHRL-TP-87-48). Brooks Air Force Base, TX: Air Force Human Resources laboratory.

Ree, M. J., & Earles, J. A. (1991). Predicting training success: Not much more than g. *Personnel Psychology, 44,* 321-332.

Ree, M. J., Earles, J. A., & Teachout, M. S. (1994). Predicting job performance; Not much more than g. *Journal of Applied Psychology, 79,* 518-524.

Rothstein, H. R. (1990). Interrater reliability of job performance ratings: Growth to asymptote with increasing opportunity to observe. *Journal of Applied Psychology, 79,* 518-524.

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., & de Fruyt, F. (2003). International validity generalization of GMA and cognitive abilities: A European community meta-analysis. *Personnel Psychology, 56,* 573-605.

Sands, B. K., Waters, B. K., & McBride, J. R. (Eds.) (1997). *Computerized adaptive testing: From inquiry to operation*. Washington, DC: American Psychological Association.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62,* 529-540.

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons learned from 26 research scenarios. *Psychological Methods, 1,199-223.*

Schmidt, F. L., Hunter, J. E., & Caplan (1981). Validity generalization results for two jobs in the petroleum industry. *Journal of Applied Psychology, 66,* 261-273.

Scullen, S. E., Bergey, P. K., & Aiman-Smith, L. (2005). Forced distribution rating systems and the improvement of workforce potential: A baseline simulation. *Personnel Psychology, 58,* 1-32.

Segal, C. (2012). Working when no one is watching: Motivation, test scores, and economic success. *Management Science, 58,* 1438-1457.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet–based tests. *Journal of Educational Measurement, 28, 237-247.*

Society for Industrial and Organizational Psychology (2003). *Principles for the validation and use of personnel selection procedures (4th edition).* Bowling Green, OH: Author.

Spearman, Charles, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3,* 271–295.

Tinsley, H. E. A., & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgements. *Journal of Counseling Psychology*, *22*, 358-376.

Visweswaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 5,* 557-574.

Wigdor, A. K., & Green, B. F. (1991). *Performance assessment in the workplace (Vol. 1).* Washington, DC: National Academy Press.

Chapter 5.
Correcting for Restriction of Test Score Range
Fritz Drasgow

Introduction

The last chapter provided a discussion about many aspects of measurement error and reliability coefficients, all having to do with the validity coefficient and its limitations. We address how to correct a validity coefficient for reliability in later chapters. This chapter focuses strictly on the correcting the validity coefficient for range restriction in test scores that results from applying a selection or classification standard (e.g., an ASVAB classification composite with cutscore) that screens out applicants below that standard. We briefly examined the effects of range restriction on the correlation coefficient in Chapter 2 for the explicit (direct) selection case. However, there are other kinds of range restriction and we address some of them in this chapter. We refer the reader to Sackett and Yang (2000) for a more complete treatment of the restriction in range issue.

Restriction in Range Situations in General

Probably the most common situation for restriction of range in test scores arises in the context of personnel selection when individuals with high scores on a selection test are admitted to college or hired for a job. Here, academic or job performance (variable *Y* to be predicted from variable *X*) can be assessed for only those individuals who are admitted or hired. Thus, whereas complete data are available for the explicit selection variable (*X)* for the total applicant population (from which, theoretically, future applicants will be selected), only partial data are available for the academic or job performance variable (*Y)*. In this case we call *Y* the *incidental* selection variable, the same as what we call a candidate *X* variable that was, for example, administered only to an already selected group of individuals based upon an operational selection or classification standard.

We describe another type of incidental selection involving just the ASVAB. Although all nine ASVAB tests are administered to military applicants, only four of the nine tests (PC, WK, MK, and AR) constituting the Armed Forces Qualification Test (AFQT) are used to determine suitability for entrance into the U.S. military services, whereas the remaining ASVAB tests (GS, MC, EI, AS, and AO) are available for military classification into occupations. If we were to use the AFQT with some cutscore to assign (classify) military selected recruits into a specific occupation subsequent to using a lower AFQT score for military selection, the AFQT would be the explicit selection variable (on a cognitive basis) for the occupation.

Now say we suspect that a different combination of ASVAB tests would be more highly predictive of training performance (the *Y* variable) than the AFQT based upon the characteristics and requirements of the occupation and the curriculum used to train for

that occupation, say a technical job that involves electronics. Such an ASVAB composite could be AR+MK+EI+GS (the Services Electronics composite). Because this composite was not used in the actual ASVAB standard, we would not expect the variance of that composite to be as diminished as the AFQT variance, given the explicit cutscore on AFQT. Because the magnitude of the correlation coefficient is dependent on test score variance, the validity of the AR+MK+EI+GS composite developed in the selected sample would be downward biased compared to the validity of the AFQT. In order to evaluate the validity of the two composites on an even playing field, we need to correct *both* validities for restriction in range of test scores and provide validity estimates for the unrestricted population.

In statistical notation, for the simplest case of one explicit predictor variable and one incidental variable (say, the criterion), we can compute the variance of *X* in the unrestricted group, S_x^2, and in the restricted (selected) group, s_x^2.[3] In contrast, we can compute the variance of the performance variable *Y*, s_y^2, in the restricted group only, because we do not have performance data for those not selected. Similarly, we can compute the correlation of *X* and *Y* in the restricted group, r_{xy}, but not in the unrestricted group, our point of interest in predictive validity studies. Thus, the variance of *Y* in the unrestricted group, S_y^2, and the correlation of *X* and *Y* in the unrestricted group, R_{xy}, are unknown and must be estimated.[4]

Restriction in range can have a very large impact on correlations (validity estimates) as observed in Table 2-4 of Chapter 2. Consequently, it is very important to understand the effects of range restriction when conducting ASVAB validation studies. As a real life example, Thorndike (1949) provided extraordinary validity results for 1,036 men in pilot training for the U.S. Army Air Force during World War II. These men were considered an experimental group and were selected without regard to their standing on a selection composite that was used under normal pilot selection circumstances (Pilot Stanine). The selection composite had a substantial correlation, R_{xy} = .64, with performance scores received in the pilot training course. Thorndike also computed the correlation for only the 136 men who *would* have qualified for pilot training under normal selection circumstances; for this selected group, the correlation of the selection composite with training school performance was only r_{xy} = .18. The importance of correcting for incidental selection as well as explicit selection can be seen in the correlations presented by Thorndike (1949), which are reproduced in Table 5-1.

[3] We use large letters for the large (i.e., unrestricted) group and small letters for the small (i.e., restricted) group.
[4] "R" in the Navy ASVAB validation/standards study context does not refer to a multiple correlation; rather it stands for the theoretical population (unrestricted) correlation (Rho, or ρ). The unrestricted validity is estimated for an ASVAB composite, consisting of a number of integer weighted ASVAB tests, as it predicts in a linear model the continuous criterion variable, final school grade in training.

Table 5-1
Thorndike's (1949) Correlations of Predictors with Success in Army Air Force Pilot Training for Total and Restricted Groups

Predictor	Total Group (*N* = 1036)	Restricted Group (*N* = 136)
Pilot Stanine (Composite Score)*	.64	.18
Mechanical Principles**	.44	.03
General Information**	.46	.20
Complex Coordination**	.40	-.03
Instrument Comprehension**	.45	.27
Arithmetic Reasoning	.27	.18
Finger Dexterity	.18	.00

*Explicit selection variable
**Components of explicit selection variable

The Pilot Stanine (Composite Score) in Table 5-1 tagged with a single asterisk was used for explicit selection and comprised the four tests tagged with a double asterisk. The other two tests without asterisks, Arithmetic Reasoning and Finger Dexterity, were given to pilot applicants as part of the Pilot selection battery but had no weight in the Pilot Stanine score. Note that one correlation in the restricted group (Complex Coordination) was negative (-.03) even though its correlation was positive and substantial in the unrestricted "Total" group (.40).

Table 5-1 shows that Instrument Comprehension had the largest correlation for the restricted group (.27) followed by General Information (.20). Complex Coordination had a negative correlation, essentially zero (-.03), and Mechanical Principles had a positive correlation, but also essentially zero (.03). These two correlations would lead one to conclude that these tests were worthless for predicting success in pilot training when in the unrestricted group from which, theoretically, future selection decisions could be made, the correlations were meaningful (.40 and .44, respectively). Most important to note in Table 5-1, the Pilot Stanine (Composite Score) has the highest validity in predicting who from the unrestricted group would be successful in Pilot training (.64). If evaluated only in the selected group, the validity would be assessed not at .64, but at only .18.

It is critical to realize that the validity estimates of interest in this example are those coefficients of the selection tests for the pilot *applicant group* from which future selection decisions will be made, not the coefficients for the selected group. Simply considering the restricted range of talent that occurs from use of a selection composite and cutscore results in a highly biased estimate of association between the selection composite and training school performance, and that bias will be downward and proportional to the reduction in variance that results from the particular cutscore.

The statistical problem in determining the applicant group validity estimates from knowledge of them only for a restricted in range group becomes somewhat of a "missing not at random" data problem. Correcting correlations and variances for this special missing data problem involves the correction for restriction of range, which has had a long history of research. Gulliksen (1950) provides a thorough treatment of bivariate, trivariate, and multivariate cases and cites work dating back to Karl Pearson (1903a, 1903b). The remainder of this chapter describes the cases of most importance to evaluating tests and composites of tests for military job classification.

The Bivariate Case: Explicit Selection on One Variable

The bivariate correction for range restriction is the simplest case and is illustrated in the Army Air Force study described in the previous section taken from Thorndike (1949). Specifically, there is explicit selection on a predictor *X* and, consequently, data are available on the criterion *Y* for only the restricted sample. Thus, we know the variance of *X* in the total group, S_x^2, as well as the selected group, s_x^2. We also know the variance of *Y* in the selected group, s_y^2, and the correlation of *X* and *Y* in the selected group, r_{xy}. Our task is to use these known quantities to determine the variance of *Y* in the total group, S_y^2, and the correlation of *X* and *Y* in the total group, R_{xy}.

We shall assume without loss of generality that all variables have been transformed to deviation scores and consequently the intercepts of regressions are all zero. Two assumptions commonly made in regression are needed to derive corrections for restriction in range: (a) linearity and (b) homoscedasticity. Stated more precisely, the first assumption is that $E(Y \mid X = x) = Bx$ for all *x*.

In the total group, the regression of *Y* on *X* is

$$\hat{Y}_T = B_{yx} X = R_{xy} \frac{S_y}{S_x} X \quad (5\text{-}1)$$

and the regression in the selected group is

$$\hat{Y}_S = b_{yx} X = r_{xy} \frac{s_y}{s_x} X . \quad (5\text{-}2)$$

Given the assumption of linearity, excluding some individuals with low *X* scores does not change the mean *Y* score for a given *X* score for those selected and so

$$B_{xy} = b_{xy} \quad (5\text{-}3)$$

and therefore

$$R_{xy}\frac{S_y}{S_x} = r_{xy}\frac{s_y}{s_x}. \quad (5\text{-}4)$$

Homoscedasticity means that the errors about the regression line have constant variance, signified by

$$S_{y\cdot x} = S_y\sqrt{1-R_{xy}^2} \quad (5\text{-}5)$$

for the total sample and

$$s_{y\cdot x} = s_y\sqrt{1-r_{xy}^2} \quad (5\text{-}6)$$

for the restricted sample. Again, excluding individuals with low X scores should not change these conditional standard deviations; therefore,

$$S_y\sqrt{1-R_{xy}^2} = s_y\sqrt{1-r_{xy}^2}. \quad (5\text{-}7)$$

Examining Equations 5-4 and 5-7 shows that we have two equations and two unknowns, S_y and R_{xy}. To solve, we first rewrite Equation 5-4 as

$$S_y = r_{xy}\frac{s_y}{s_x}\frac{S_x}{R_{xy}}, \quad (5\text{-}8)$$

and then substitute into Equation 7,

$$r_{xy}\frac{s_y}{s_x}\frac{S_x}{R_{xy}}\sqrt{1-R_{xy}^2} = s_y\sqrt{1-r_{xy}^2}. \quad (5\text{-}9)$$

Squaring both sides and moving known quantities to the right side,

$$\frac{1-R_{xy}^2}{R_{xy}^2} = \frac{s_x^2}{S_x^2}\frac{1-r_{xy}^2}{r_{xy}^2} \quad (5\text{-}10)$$

and therefore

$$\frac{1}{R_{xy}^2} - 1 = \frac{s_x^2}{S_x^2}\left(\frac{1}{r_{xy}^2} - 1\right). \quad (5\text{-}11)$$

Adding one to both sides and taking the reciprocals,

$$R_{xy}^2 = \frac{1}{1 + \frac{s_x^2}{S_x^2}\left(\frac{1}{r_{xy}^2} - 1\right)}, \quad (5\text{-}12)$$

and

$$R_{xy} = \frac{1}{\sqrt{1 + \frac{s_x^2}{S_x^2}\left(\frac{1}{r_{xy}^2} - 1\right)}}, \quad (5\text{-}13)$$

which provides the formula to convert the known correlation from the restricted sample, r_{xy}, to an estimate of the unknown correlation R_{xy} in the total group, the validity of interest when validating tests for selection purposes. Also, given this value of R_{xy}, we can substitute into Equation 5-8 to obtain an estimate of S_y.

The Trivariate Case: Implicit Selection on a Third Variable

In the trivariate case, we have a predictor X, criterion Y, and an additional predictor Z that we wish to study. There is direct selection on X, which reduces its correlation with Y in the selected group as described in the previous section. The additional predictor Z is typically not used to make selection decisions, so there is no *explicit* selection on Z. However, ordinarily Z and X are correlated, so selecting on X has the effect of reducing the variance of Z. This situation is called *indirect* or *incidental* selection on Z.

Of critical importance in the situation of incidental selection is that the practitioner understands the experimental design and at what point Z was administered. In a *predictive* validity study, applicants are administered the selection test X and the new predictor Z *at the same time*. Explicit selection on X occurs and selectees report to, say, training. Training performance Y scores are observed along with X and Z scores for all those who trained. In a *concurrent* validity study, applicants are administered only the selection test X. As before, there is explicit selection on X and selectees report to training. Z is administered to all selectees who report to training and, as before, X and Z scores are observed for all those who trained. The key distinction between the predictive and concurrent designs is whether the variance of Z is observed in the total group or not (i.e., is S_z^2 known).

In the predictive validity study, all applicants are given tests X and Z, so S_x^2 and S_z^2 are known. Performance is observed later, so s_y^2 is known but not S_y^2. The correlations of X and Z with performance Y are observed in the selected group, so r_{xy} and r_{zy} are known, but the correlations of the two predictors with the criterion in the total group, R_{xy} and R_{zy}, must be estimated.

In the predictive validity design, we need to assume that the regression of Y on Z is linear and homoscedastic as well as Y on X. Given the assumption of linearity, selection on X does not affect the conditional means of Y given Z and we have

$$B_{yz} = R_{zy}\frac{S_y}{S_z} = r_{zy}\frac{s_y}{s_z} = b_{yz}. \quad (5\text{-}14)$$

The homoscedasticity assumption means

$$S_{y\cdot z} = S_y\sqrt{1-R_{zy}^2} = s_y\sqrt{1-r_{zy}^2} = s_{y\cdot z}. \quad (5\text{-}15)$$

As in the case of direct selection on X, we have two equations and two unknowns (S_y and R_{zy}); note that S_z, s_z, s_y, and r_{zy} are known. Solving Equations 5-14 and 5-15 gives

$$S_y = r_{zy}\frac{s_y}{s_z}\frac{S_z}{R_{zy}} \quad (5\text{-}16)$$

and

$$R_{zy} = \frac{1}{\sqrt{1+\frac{s_z^2}{S_z^2}\left(\frac{1}{r_{zy}^2}-1\right)}}, \quad (5\text{-}17)$$

which have exactly the same form as Equations 5-8 and 5-13. Thus, the fact that there is direct selection on X and incidental selection on Z is immaterial. What is important is that S_x^2 and S_z^2 are known.

In a concurrent validity study, S_z^2 is not known and so we cannot simply use Equation 5-17 to estimate R_{zy}. In this case, S_y, S_z, and R_{zy}, are unknowns, so we have two equations (Equations 5-14 and 5-15) and three unknowns. To obtain a solvable set of equations, we need the additional assumption that the partial correlation between Z and Y, holding X constant, is the same in the restricted and total groups. Specifically,

$$R_{zy\cdot x} = \frac{R_{zy} - R_{xz}R_{xy}}{\sqrt{1-R_{xz}^2}\sqrt{1-R_{xy}^2}} = r_{zy\cdot x} = \frac{r_{zy} - r_{xz}r_{xy}}{\sqrt{1-r_{xz}^2}\sqrt{1-r_{xy}^2}}. \quad (5\text{-}18)$$

Equations 5-14, 5-15, and 5-18 contain the three unknowns, and hence can be solved (see p. 149 in Gulliksen [1950] for the algebra). The solution for the correlation of *Z* and *Y* for the total group is

$$R_{zy} = \frac{r_{zy} - r_{xz}r_{xy} + r_{xz}r_{xy}\dfrac{S_x^2}{s_x^2}}{\sqrt{\left[1-r_{xz}^2+r_{xz}^2\dfrac{S_x^2}{s_x^2}\right]}\sqrt{\left[1-r_{xy}^2+r_{xy}^2\dfrac{S_x^2}{s_x^2}\right]}}. \quad (5\text{-}19)$$

The Multivariate Case

Sackett and Yang (2000), in their article that addresses expanded types of restriction in range, cited both Aitken (1934) and Lawley (1943) as further developing the published Pearson (1903) correction formulas for the multivariate case. Hunter, Schmidt, and Le (2006) point out that the military is in the favorable position of applying the multivariate correction formulas when evaluating candidate ASVAB composites that are subject to incidental selection because all military applicants are required to take the full ASVAB (and there is an ASVAB normative youth population that can serve as the unrestricted population in the corrections).

Gulliksen (1950) discussed several cases of multivariate selection (e.g., the variances of the incidental selection variables for the unrestricted population are known or unknown). As in the case of the predictive versus concurrent study design, the key to the correction formulas is whether the variances of the predictors are known in the population. In the case of the ASVAB, the variances of all the ASVAB tests, and the explicit selection composite formed from these tests, are known for both the unrestricted population (which for the Navy, is the ASVAB normative population) and the restricted "selected" group, but the variance of the criterion variable is known only for the selected group. In this case, all of the ASVAB tests, which provide more potentially relevant information about an applicant, can be treated mathematically as explicit selection variables when in fact a single composite of ASVAB tests used for explicit selection (more information about the applicant is obtained through use of all ASVAB tests).

We adopt some matrix algebra notation for the multivariate case (generally cited by the Navy as Lawley [1943]). Let $\boldsymbol{C}_{xx}$ and $\mathbf{c}_{xx}$ denote the variance-covariance matrices of the predictors in the total and restricted groups, $\boldsymbol{C}_{xy}$ and $\mathbf{c}_{xy}$ denote the vector of covariances of the predictors with the criterion in the total and restricted groups, and, as before, S_y^2 and s_y^2 denote the variance of the criterion in the total and restricted groups.

As in the univariate case, we assume that the regression of the criterion on the predictors is unaffected by selection,

$$\mathbf{B}_{xy} = \mathbf{C}_{xx}^{-1}\mathbf{C}_{xy} = \mathbf{b}_{xy} = \mathbf{c}_{xx}^{-1}\mathbf{c}_{xy}. \quad (5\text{-}20)$$

From Equation 5-20 it is apparent that

$$\mathbf{C}_{xy} = \mathbf{C}_{xx}\mathbf{c}_{xx}^{-1}\mathbf{c}_{xy} = \mathbf{C}_{xx}\mathbf{b}_{xy}, \quad (5\text{-}21)$$

so that the covariances of the predictors can be computed from known quantities ($\mathbf{C}_{xx}$, $\mathbf{c}_{xx}$, and $\mathbf{c}_{xy}$). The standard deviations of the predictors are known (they are the square roots of the diagonal entries in $\mathbf{C}_{xx}$) and so all that remains is determining the variance of the criterion in the total group.

Using the theory of linear transformations, the variance of the errors $E = Y - \hat{Y}$ is

$$\begin{aligned} Var(Y-\hat{Y}) &= Var(Y) + Var(\hat{Y}) - 2Cov(Y,\hat{Y}) \\ &= S_y^2 + Var(\mathbf{X}'\mathbf{B}_{xy}) - 2Cov(Y,\mathbf{X}'\mathbf{B}_{xy}) \\ &= S_y^2 + \mathbf{B}_{xy}'\mathbf{C}_{xx}\mathbf{B}_{xy} - 2\mathbf{C}_{xy}\mathbf{B}_{xy} \\ &= S_y^2 + \mathbf{C}_{xy}'\mathbf{C}_{xy}^{-1}\mathbf{C}_{xx}\mathbf{B}_{xy} - 2\mathbf{C}_{xy}'\mathbf{B}_{xy} \\ &= S_y^2 + \mathbf{C}_{xy}'\mathbf{B}_{xy} - 2\mathbf{C}_{xy}'\mathbf{B}_{xy} \\ &= S_y^2 - \mathbf{C}_{xy}'\mathbf{B}_{xy} \end{aligned} \quad (5\text{-}22)$$

for the total group and for the restricted group,

$$Var(Y-\hat{Y}) = s_y^2 - \mathbf{c}_{xy}'\mathbf{b}_{xy}. \quad (5\text{-}23)$$

If the errors are homoscedastic,

$$S_y^2 - \mathbf{C}_{xy}'\mathbf{B}_{xy} = s_y^2 - \mathbf{c}_{xy}'\mathbf{b}_{xy} \quad (5\text{-}24)$$

so that,

$$S_y^2 = s_y^2 - \mathbf{c}_{xy}'\mathbf{b}_{xy} + \mathbf{C}_{xy}'\mathbf{B}_{xy}. \quad (5\text{-}25)$$

Using Equation 5-20 yields

$$S_y^2 = s_y^2 + (\mathbf{C}_{xy}' - \mathbf{c}_{xy}')\mathbf{b}_{xy}. \quad (5\text{-}26)$$

Define the diagonal matrix $\mathbf{D}_x$ as containing the variances of the predictors,

$$\mathbf{D}_x = Diag(\mathbf{C}_{xx}). \quad (5\text{-}27)$$

Then the correlations of the predictors with the criterion in the total group are

$$\mathbf{R}_{xy} = \frac{1}{S_y}\mathbf{D}_x^{-1/2}\mathbf{C}_{xy}. \quad (5\text{-}28)$$

Concluding Remarks

There are several cases of explicit and incidental selection situations that should be considered when correcting validity coefficients for restriction in range. However, the Navy, in validating the operational and candidate replacement ASVAB composites for a specific occupation classification standard, applies the multivariate correction for range restriction that treats all nine of the ASVAB tests as explicit selection variables. The multivariate method in many cases has been found to give more accurate estimates of population validity coefficients than the univariate method and also addresses the incidental selection situations involving ASVAB tests that are not used in the operational selection composite. A number of issues regarding this procedure do exist and are discussed in later chapters (see for example, negative range corrected validity coefficients discussed in Chapter 11). We also refer the reader to Dunbar and Linn (1991) for more about the restriction in range topic in a military context. The next chapter addresses the joint correction for range restriction and measurement error.

Chapter 5. References

Aitken, A. C. (1934). Note on selection from a multivariate normal population. *Proceedings of the Edinburgh Mathematical Society, 4,* 106-110.

Dunbar, S. B., & Linn, R. L. (1991). Range restriction adjustments. In A. K. Wigdor & B. F. Green (Eds.), *Performance assessment for the workplace*, *Vol II - Technical issues* (pp. 127-157). Washington, DC: National Academy Press.

Gulliksen, H. (1950). *Theory of mental tests*. NY: Wiley.

Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology, 91,* 594-612.

Lawley, D. (1943). A note on Karl Pearson's selection formula. *Royal Society of Edinburgh, Proceedings, Section A, 62,* 28-30.

Pearson, K. (1903a). Mathematical contributions to the theory of evolution - XI. On the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions of the Royal Society, London, Series A,* 200, 1-66.

Pearson, K. (1903b). On a general theory of the method of false position. *Philosophical Magazine, 4,* 658-668, 6th series.

Ree, M. J., Carretta, T. R., Earles, J. A., & Albert, W. (1994). Sign changes when correcting for range restriction: A note on Pearson's and Lawley's selection formulas. *Journal of Applied Psychology, 79*, 298-301.

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology*, *85*, 112-118.

Thorndike, R. L. (1949). *Personnel selection*. NY: Wiley.

Chapter 6.
Joint Corrections for Measurement Error and Range Restriction
Sarah A. Hezlett

Introduction

Both restriction in range of test scores and measurement error attenuate observed validity coefficients, making joint corrections appealing. Historically, however, there has been confusion about how to correct for both range restriction and unreliability, or even if such corrections are appropriate. Recent theoretical work has clarified that the nature of the range restriction drives the appropriate procedures to follow when making joint corrections (Hunter, Schmidt, and Le, 2006; Mendoza & Mumford, 1987; Stauffer & Mendoza, 2001). This chapter provides some background on the joint corrections.

Background

The formulas for correcting for correlation attenuation due to measurement error (attenuation) and range restriction were derived independently, creating some uncertainty about how joint corrections should be made (Stauffer & Mendoza, 2001). Standard practices have ranged from stern cautions about their use to almost nonchalant applications of accepted rules of thumb. For example, *Standards for Educational and Psychological Tests* (APA, 1974, revised in 1999) advised that validity coefficients corrected for both measurement error and range restriction should only be used to guide further research, but a psychometric rationale was not included to support this warning (Bobko, 1983; Schmidt, Hunter, Pearlman, & Hirsh, 1985). More recent validation guidelines (e.g., see Chapter 7 of the Introductory Manual and SIOP, 2003) encourage such corrections.

At other times, the methods of combining the two types of corrections have almost appeared to be taken lightly. The corrections for range restriction and measurement error have been treated as if they affect the validity coefficient separately and can be combined linearly (Hunter et al., 2006; Mendoza & Mumford, 1987). The order in which corrections have been made traditionally has been driven by the nature of the available data. Specifically, whether or not the estimate of reliability was based on range restricted data has been used as the factor determining the sequence of corrections (Stauffer & Mendoza, 2001) (explained more fully in the next chapter).

Restricting the range of data not only affects the magnitude of the observed validity coefficient (as described in Chapter 5); it also attenuates reliability estimates (the correlation of true scores with observed scores). For example, in many validation studies, criterion data (scores) are only available for job incumbents who have been hired on the basis of their scores on a selection measure. The range of incumbent scores on the criterion is restricted in comparison to what the range of scores would have been if all applicants, rather than just those who performed well on the selection instrument,

had been hired and proceeded to perform. Assuming reliability is constant over the total range of criterion scores (which may not be the case), the magnitude of a reliability estimate for the criterion data collected from job incumbents will be biased downward from that which would be obtained in an unrestricted sample of job applicants (theoretically having performance scores).

Conventionally, if the estimate of reliability was known for the unrestricted group (not affected by range restriction), the validity coefficient was first corrected for range restriction and then corrected for measurement error. On the other hand, if the estimate of the reliability was curtailed through range restriction, the validity coefficient was first corrected for unreliability and then corrected for range restriction (Stauffer & Mendoza, 2001). In some situations, however, using this rule of thumb will yield inappropriate results (Stauffer & Mendoza).

On issue with the rule of thumb is that it has been demonstrated that unreliability and range restriction interact, affecting how range restriction is defined statistically (Hunter et al., 2006; Mendoza & Mumford, 1987). Consequently, the nature of the range restriction, rather than whether or not the reliability estimate is affected by range restriction, should determine how joint corrections for measurement error and range restriction are made (Hunter et al.; Stauffer & Mendoza, 2001). The appropriate steps and formulas to use in correcting jointly for unreliability and range restriction depends upon (a) the nature of the range restriction, (b) the type of data available, and (c) the objectives of the research (i.e., whether correcting for measurement error in the predictor is appropriate – not considered so in operationally focused ASVAB validation/standards studies).

In the next section, we review derivations of the correction for measurement error, followed by the corrections for both measurement error and range restriction.

Correcting Validity Coefficients for Measurement Error

A mathematical formula specifying the relation of the correlation between observed measures with the correlation between true scores was discussed by Spearman in 1904, making it one of the earliest applications of classical, or true-score, test theory (Charles, 2005; Muchinsky, 1996). According to this formula, the hypothetical correlation between observed scores on two measures (ρ_{xy}) is a function of the correlation between the variables the measures are designed to assess (i.e., their true scores, T_x and T_y) and the reliabilities of the measures (ρ_{xx} and ρ_{yy}) (Charles):

$$\rho_{xy} = \rho_{T_x T_y} \sqrt{\rho_{xx}} \sqrt{\rho_{yy}} \; . \quad (6\text{-}1)$$

Equation 6-1 can be algebraically solved to obtain ρ_{TxTy}, the correlation between the true scores of X and Y, that is, an estimate of an observed correlation corrected for measurement error in both the predictor and criterion having estimates of the reliability of both measures. Equation 2-16 in Chapter 2 is that algebraic solution and is restated here as:

$$\rho_{T_x T_y} = \frac{\rho_{xy}}{\sqrt{\rho_{xx}}\sqrt{\rho_{yy}}} = \rho_{x_c y_c}. \quad (6\text{-}2)$$

Note that the equation of the correlation between true scores and the corrected value rests on the assumption that the errors are random (Charles, 2005) and not subject to sampling issues. The practitioner's use of the correction for attenuation presented in Equation 6-2 requires the substitution of observed sample statistics presented as Equation 6-3:

$$r_{x_c y_c} = \frac{r_{xy}}{\sqrt{r_{xx}}\sqrt{r_{yy}}}. \quad (6\text{-}3)$$

Although many researchers have discussed an observed validity coefficient corrected for measurement error as if it *were* a population value, we must remember that the correction is based on sample statistics and is thus merely an estimate of the population parameter (Charles, 2005; Muchinsky, 1996) and one that rests on the assumptions of classical test theory (CTT). (CAT-ASVAB reliabilities are now based upon Item Response Theory (IRT), which has a different set of assumptions and a different way in quantifying measurement precision, briefly discussed later.)

The corrected validity coefficient for measurement error is generally a less biased estimate of the population parameter than is the uncorrected validity coefficient (SIOP, 2003). However there can be overestimates or underestimates of the population validity coefficient, as highlighted by the fact that the value of a corrected validity coefficient occasionally can be greater than 1. That is, correction for attenuation sometimes yields values for validity coefficients that are not theoretically possible. Debate over the possible causes of this phenomenon erupted soon after Spearman's (1904) work on the correction for attenuation and sampling error; errors in estimating reliability remain viable explanations (Charles, 2005; Muchinsky, 1996).

In practice, the following correction formula is generally used in industry for hiring decisions, as it is widely agreed that only error in measuring the criterion should be corrected, not error in measuring predictors (SIOP, 2003).

$$r_{xy_c} = \frac{r_{xy}}{\sqrt{r_{yy}}}. \qquad (6\text{-}4)$$

In industry, hiring decisions must be based on the fallible scores collected with the predictor (Muchinsky, 1996), making a validity coefficient corrected for the reliability of the predictor a poor estimate of the predictor's operational validity. We recognize that the Navy does not correct for reliability for the training criterion used to validate the ASVAB in the operational occupational classification context, which will be discussed in Chapter 17 about setting ASVAB cutscores. When validity coefficients are corrected for measurement error, both the corrected and uncorrected values should be reported (SIOP, 2003).

The Joint Correction for Direct Range Restriction

First, the case of direct range restriction due to explicit selection on an observed variable (e.g., scores on the ASVAB) is considered, and then the case of incidental selection. In both cases, emphasis is placed on outlining the steps to follow when correcting for range restriction and unreliability only in the criterion. Procedures that also incorporate corrections for measurement error in the predictor are mentioned, but treated more in depth in the chapter that follows. The correction for measurement error in the predictor is typically applied for research purposes (e.g., evaluating a potential new addition to the ASVAB) but are rarely used to inform decisions about operational selection systems, our focus in this chapter. We note that the addition of corrections for measurement error in the predictor changes the appropriate steps to follow, at times in substantial ways.

For many jobs, applicants take a selection test. The *x* scores of the unrestricted sample of applicants are used to make hiring decisions. The resulting pool of job incumbents has a restricted set of scores on the selection test (*x*). As noted in Chapter 5, the standard deviation of scores in the unrestricted applicant population (S_x) differs from the standard deviation of scores in the restricted job incumbent sample (s_x). In almost all cases involving educational or personnel selection, criterion data (such as training or job performance) are only available in the restricted sample (*y*). Thus, estimates of criterion reliability typically are based on data collected from the restricted sample of job incumbents (r_{yy}), rather than on the unrestricted sample of applicants (r_{YY}).

In this situation where criterion reliabilities are only available in the restricted in range sample, and the interest is in the validity of an explicit selector, several different approaches may be used to correct observed validity coefficients for both measurement error in the criterion and range restriction (Bobko, 1983; Hunter et al., 2006; Lee, Miller, & Graham, 1982). A three-step procedure was developed by Schmidt, Hunter, and Urry (1976) that consists of (a) correcting the observed criterion reliability for range restriction, (b) correcting the observed validity for range restriction, and (c) using the range corrected reliability coefficient to further correct the validity coefficient (that has been corrected for range restriction). A two-step correction procedure involves (a) correcting the observed validity coefficient for criterion unreliability using the observed reliability coefficient and (b) correcting the resulting validity coefficient that has been corrected for unreliability for restriction of range (Lee et al.). The two procedures are mathematically identical (Bobko) and may be combined in a single step utilizing one formula (Bobko; Hunter et al.). Research has demonstrated that these corrections for both range restriction and measurement error in the criterion yield estimates of the correlation between the variables of interest that are less biased than the uncorrected, observed correlation(Lee et al.; Bobko).

More formally, when correcting an observed validity for direct range restriction and measurement error in the criterion, the first step is to correct the correlation between the predictor and criterion in the restricted sample (i.e., r_{xy} for the job incumbent sample) using an estimate of the reliability of the criterion in the restricted sample (r_{yy})

(Hunter et al., 2006). As the following formula shows, this yields an estimate of the correlation between the predictor (x) and the estimated true score on the criterion (y_c) in the restricted population (Hunter et al.):

$$r_{xy_c} = \frac{r_{xy}}{\sqrt{r_{yy}}} . \quad (6\text{-}5)$$

The next step is to correct the correlation for range restriction on the predictor, which results in an estimate of the correlation between scores on the predictor and the estimated true scores on the criterion for the unrestricted population. This corrected validity coefficient (now using R as the estimated population validity) is considered to be an estimate of the "operational validity" of the predictor (Hunter et al., 2006):

$$R_{XY_c} = \frac{r_{xy_c} U_X}{\sqrt{1 + (U_X^2 - 1) r_{xy_c}}} , \quad (6\text{-}6)$$

where, for simplification purposes, $U_x = S_x/s_x$. The operational validity also can be obtained in a single step using the following equation (Bobko, 1983; Hunter et al.). Note that the inputs on the right-hand side of the equation are based on the restricted sample of job incumbents; that is, they are the observed validity coefficient and the estimate of criterion reliability based on the restricted sample. The only piece of data from the unrestricted data set (i.e., the applicant population) that is utilized is the standard deviation S_x in U_x:

$$R_{XY_c} = \frac{r_{xy} U_X}{\sqrt{r_{yy} + U_X^2 r_{xy}^2 - r_{xy}^2}} , \quad (6\text{-}7)$$

For most purposes, the correction process would stop with either the use of Equation 6-6 (if the multiple step procedures were followed) or Equation 6-7. However, for some theory based research, an additional correction for measurement error in the predictor can be made to obtain the correlation between estimated predictor *true* scores and estimated criterion true scores in the unrestricted population, which merely involves dividing the operational validity from Equation 6-7, by the square root of the estimate of the reliability of the predictor in the unrestricted (applicant) population (Hunter et al., 2006):

$$R_{X_c Y_c} = \frac{R_{XY_c}}{\sqrt{R_{XX}}} . \quad (6\text{-}8)$$

Hunter et al. (2006) provide a formula for obtaining the validity coefficient corrected for range restriction, criterion unreliability, and predictor unreliability in a single step. Alternate approaches may be used if an estimate of the criterion's reliability is only available for the unrestricted (applicant) population (Hunter et al.). It should be noted that the correction procedures are best applied in large samples (Mendoza & Mumford, 1987) and that we should remember that just as standard errors are important to consider in the estimation of basic kinds of statistics, they are also important for estimating the standard errors of the joint corrections (e.g., Fife, Mendoza, & Terry, 2012).

The Joint Correction for Indirect Range Restriction

Correcting for range restriction and measurement error is a more complex process when there is indirect range restriction. Within selection contexts, a common situation that illustrates indirect range restriction is a concurrent validation study (Hunter et al., 1986). In a concurrent validation study, job incumbents are measured on both a "potential" selection measure (X) and a criterion measure (Y). The incumbents were not hired on the basis of their scores on the potential selection measure but on Z, which proceeded measurement on both the potential predictor and the criterion (Hunter et al., 2006) (notice we are using Z now to designate the explicit selector, not X). If the original method of selecting the job incumbents (Z) correlated with the potential selection measure, the hiring of the job incumbents was reflected in their true scores (T) on the potential predictor X.

In essence, in the scenario just described, selection has technically been made on the basis of the latent ability (T) assessed by X (Hunter et al., 2006; Mendoza & Mumford, 1987). Correcting for criterion measurement error and indirect range restriction may be accomplished in a multiple step procedure (Hunter et al.); however, the steps needed will vary depending upon the reliability estimates available for the potential predictor, X. We note that it is possible to estimate the reliability of X in the restricted group from the reliability of X in the unrestricted group and vice versa (Hunter et al.); therefore, the steps executed will depend on what reliabilities need to be computed (shown shortly).

First, the observed correlation between the potential predictor and criterion in the restricted (i.e., job incumbent) sample (r_{xy}) is corrected using an estimate of the reliability of the criterion (r_{yy}) in the restricted sample (Hunter et al., 2006) as was done in the direct selection case (Equation 6-1) repeated here as Equation 6-9.

$$r_{xy_c} = \frac{r_{xy}}{\sqrt{r_{yy}}}. \quad \text{(6-9)}$$

As Equation 6-9 shows (again), the correction yields an estimate of the correlation between the potential predictor in the incumbent sample and the estimated true score on the criterion in this sample (Hunter et al., 2006).

Next, if the reliability of X (r_{xx}) in the restricted (incumbent) sample is not known, it must be estimated from the reliability of X in the unrestricted (applicant) population (R_{XX}) assuming it has been reported. As the following equation illustrates, and as Hunter et al. (2006) note, "…the incumbent reliability of the independent variable may be considerably lower than the applicant reliability" (p. 602):

$$r_{xx} = 1 - U_X^2(1 - R_{xx}). \quad (6\text{-}10)$$

The step involving Equation 6-10 may be skipped if there is already an estimate of the restricted reliability (Hunter et al.) (e.g., from test-retest administration of the measure during the predictor developmental stage).

The third step in the procedure involves correcting r_{xy} for measurement error in X found estimated for the restricted in range population (Hunter et al., 2006). Note that this is a crucial place where the sequencing of the steps for direct and indirect range restriction diverges. In the case of indirect range restriction, the correlation between the predictor scores and the criterion true scores for the restricted group (r_{xy_c}) are corrected for the unreliability of the potential predictor in the restricted population (r_{xx}), yielding an estimate of the correlation between predictor and criterion true scores in the restricted sample:

$$r_{x_c y_c} = \frac{r_{xy_c}}{\sqrt{r_{xx}}}. \quad (6\text{-}11)$$

The corrected correlation ($r_{x_c y_c}$) in Equation 6-11 is not estimated in the steps to correct for direct range restriction.

The fourth step in making corrections for measurement error and indirect range restriction is to estimate the reliability of the predictor, X, in the unrestricted population R_{XX}, if it is not known (Hunter et al., 2006), by

$$R_{XX} = 1 - u_x^2(1 - r_{xx}). \quad (6\text{-}12)$$

Note that $u_X = s_x / S_X = 1/U_X$ where U_x was defined earlier and which shows that R_{XX} and r_{xx} can be calculated from each other. This step can be skipped if an estimate of R_{XX} is available for the unrestricted population (as it is for an ASVAB composite when it is used as the explicit selection variable and other ASVAB composites are subject to incidental selection effects). The fifth step involves estimating the range restriction which has occurred on the latent trait or ability (T) that is assessed by X shown as

$$u_T = \sqrt{[u_X^2 - (1 - R_{XX})]/R_{XX}}. \quad (6\text{-}13)$$

In the sixth step, this estimate of u_T is used to correct for the effect of indirect range restriction, yielding an estimate of the correlation between true scores on the predictor and criterion in the unrestricted population (Hunter et al., 2006). Note again that $U_T = 1/u_T$. Thus, the outputs of the previous (fifth) step and the third step in the procedure are used as inputs to this step yielding

$$R_{X_c Y_c} = \frac{r_{x_c y_c} U_T}{\sqrt{1 + (U_T^2 - 1) r_{x_c y_c}^2}}. \quad (6\text{-}14)$$

$R_{X_c Y_c}$ is an estimate of the validity of the predictor corrected for indirect range restriction, predictor unreliability, and criterion unreliability. For some research purposes, this is the estimate of interest. However, as we have noted, for most applied decisions about the use of *X*, an additional step is needed to estimate the operational validity of *X*. In essence this step re-introduces measurement error and yields an estimate of the correlation between the predictor scores and true scores on the criterion in the unrestricted group (Hunter et al., 2006).

$$R_{XY_c} = R_{X_c Y_c} \sqrt{R_{XX}}. \quad (6\text{-}15)$$

We note here that Mayberry and Wright (1992) (Table 4.2 in Chapter 4) used Equation 6-12 (a form also shown in Lord & Novick, 1968, Equation 6.2.1) to estimate the unrestricted reliabilities of their Job Performance Measurement (JPM) project *criterion* measures (Table 4.2 in Chapter 4). Substituting *Y* for *X* and recouping the components of the *U* ratio we show:

$$\hat{\rho}_{YY} = 1 - \frac{s_y^2}{S_Y^2}(1 - r_{yy}).$$

Mayberry and Write obtained an estimate of the unrestricted *Y* standard deviation from the same procedure used to perform the multivariate correction for range restriction (Chapter 5) on ASVAB restricted in range validity coefficients (e.g., Lawley, 1943). The same method was also used in the Enhanced Computer Administered Test (ECAT) battery project (Wolfe, Alderton, Larson, & Held, 1995) that involved not only potential new ASVAB tests, but more realistic performance based criterion measures derived from the schoolhouse setting (Kieckhaefer et al., 1992).

Concluding Remarks

This chapter was intended to inform ASVAB validation/standards researchers about the complicated validity corrections that are considered in the psychometric and industrial-organizational, education research and operational settings. Combining corrections for range restriction and unreliability is a complex process. Selecting the

appropriate correction procedures requires careful consideration of the nature of the range restriction, the objectives of the research, and the available data. Research has demonstrated that, in general, joint corrections yield less biased estimates of the relation of interest (Bobko, 1983; Hunter et al., 2006) and so these corrections have been recommended (SIOP, 2003). We note, however, that sampling error is not accounted for in any of this chapter's correction formulas and as always the researcher should be mindful of small samples and any other factors that may result in spurious findings.

The next chapter provides a further discussion of the joint corrections for reliability and restriction in range from a slightly different perspective to reinforce important principles that are grounded in classical measurement theory.

Chapter 6. References

American Educational Research Association/American Psychological Association/ National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bobko, P. (1983). An analysis of correlations corrected for attenuation and range restriction. *Journal of Applied Psychology, 68,* 584-589.

Charles, E. P. (2005). The correction for attenuation due to measurement error: Clarifying concepts and creating confidence sets. *Psychological Methods, 10,* 206-226.

Fife, D. A., Mendoza, J. L., & Terry, T. (2012). The assessment of reliability under range restriction: A comparison of α, ω, and test-retest reliability for dichotomous data. *Educational and Psychological Measurement, 19,* 862-888.

Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology, 91,* 594-612.

Keickhaefer, W. F., Ward, D. G., Kusulas, J. W., Cole, D. R., Rupp, L. M., & May, M. H. (1992, September). *Criterion development for 18 technical training schools in the Navy, Army, and Air Force* (Contract # N66001-90-D-9502, Delivery Order 7J08). San Diego, CA: Navy Personnel Research and Development Center.

Lawley, D. (1943). A note on Karl Pearson's selection formula. *Royal Society of Edinburgh, Proceedings, Section A, 62,* 28-30.

Lee, R., Miller, K. J., & Graham, W. K. (1982). Corrections for restriction of range and attenuation in criterion-related validation studies. *Journal of Applied Psychology, 67,* 637-639.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Mayberry, P. W., & Wright, W. H. (1992). *Reliability of mechanical maintenance performance measures* (CRM 91-246). Alexandria, VA: Center for Naval Analyses.

Mendoza, J. L., & Mumford, M. (1987). Corrections for attenuation and range restriction on the predictor. *Journal of Educational Statistics, 12,* 282-293.

Muchinsky, P. M. (1996). The correction for attenuation. *Educational and Psychological Measurement, 56,* 63-75.

Schmidt, F. L., Hunter, J. E., Pearlman, K., & Hirsh (1985). Forty questions about validity generalization and meta-analysis. *Personnel Psychology, 38,* 697-798.

Schmidt, F. L., Hunter, J. E., & Urry, V. W. (1976). Statistical power in criterion-related validity studies. *Journal of Applied Psychology, 61,* 473-485.

Society for Industrial and Organizational Psychology (2003). *Principles for the validation and use of personnel selection procedures*. Bowling Green, OH: SIOP.

Spearman, C. (1904). The proof and measurement association between two things. *The American Journal of Psychology, 15,* 72-101.

Stauffer, J. M., & Mendoza, J. L. (2001). The proper sequence for correcting correlation coefficients for range restriction and unreliability. *Psychometrika, 66,* 63-66.

Wolfe, J. H., Alderton, D. L., Larson, G. E., & Held, J. D. (1995). *Incremental validity of Enhanced Computer Administered Testing (ECAT)* (NPRDC –TN-96-6). San Diego, CA: Navy Personnel Research and Development Center.

Chapter 7.
More on Joint Corrections
Jorge L. Mendoza

Introduction

As we saw in the last chapter, correcting for both unreliability in measures and restriction in range due to selection effects is a bit more complicated than either correction alone. Hunter, Schmidt, and Le (2006) refer to measurement error as a "simple artifact" as would be a dichotomized quantitative variable. Simple artifacts combine in a linear fashion without regard to order so that their correction formulas can be rolled up into one "global artifact" term that is then used, for example, to correct a correlation. Restriction in range, however, introduces complications and interactions, thereby referred to as "a complex artifact". Not only must one consider the degree of restriction in range imposed by the cutscore on a selection instrument, but also the magnitude of the population validity coefficient, which is in itself what the researcher is trying to determine. In this chapter we attempt to somewhat simplify the concepts of the joint correction with strict adherence to classical test theory (CTT) assumptions. We specifically address the problem that arises when direct selection results in a violation of the assumption that true and error scores are not correlated. Some of the concepts and formulas from the last chapter are represented here from a slightly different perspective but should on balance be more helpful for instantiating principles than a distraction. Note that in this chapter we use the asterisk to indicate the range restricted statistics.

The Joint Correction Paradigm

As noted in the previous chapter, in correcting for both range restriction and unreliability, the sequence has usually been a matter of the estimates of reliability that are available to the researcher. If we assume that the researcher is attempting to go from the restricted observed correlations, $r^{*}_{x,z,y}$ to the unrestricted "true" correlations, $r_{tx,tz,ty}$, depending on whether we have restricted or unrestricted reliability estimates, the sequence of corrections differs. When the reliability estimate is restricted, we follow the sequence of correcting for unreliability first and then for range restriction, as follows:

$$r^{*}_{x,z,y} \rightarrow r^{*}_{tx,tz,ty} \rightarrow r_{tx,tz,ty} .$$

When the reliability estimate is unrestricted, we correct first for range restriction, then for unreliability as follows:

$$r^{*}_{x,z,y} \rightarrow r_{x,z,y} \rightarrow r_{tx,tz,ty} .$$

When the reliability estimates are mixed (e.g., an unrestricted reliability for variable x, and a restricted reliability for variable y from the restricted sample) we must bring them all to the same level before correcting (discussed in the last chapter).

The second sequence, correcting first for range restriction and last for reliability, is preferred because biased estimates of reliability have been observed for the explicit selection variable (say x) under direct (explicit) range restriction (Fife, Mendoza, Terry, 2012). This matter is discussed later in the chapter.

In principle, the joint correction is simple, and it follows the same pattern regardless of the number of variables involved. To illustrate, consider our preferred correction sequence in matrix format for a three-variable situation that has been observed in the selected sample. (The reader may want to refer back to Chapter 5 for the matrix form of the multivariate correction for range restriction.) We show the matrix sequence as,

$$r^{*}_{x,z,y} = \begin{pmatrix} 1 & r^{*}_{xz} & r^{*}_{xy} \\ & 1 & r^{*}_{zy} \\ & & 1 \end{pmatrix} \longrightarrow r_{x,z,y} = \begin{pmatrix} 1 & r_{xz} & r_{xy} \\ & 1 & r_{zy} \\ & & 1 \end{pmatrix} \longrightarrow r_{tx,tz,ty} = \begin{pmatrix} 1 & r_{txtz} & r_{txty} \\ & 1 & r_{tzty} \\ & & 1 \end{pmatrix}$$

The matrix on the left contains the restricted correlations among x, z, and y where (a) x was used as an explicit (direct) selection variable, (b) y is the performance measure taken sometime after selection and incidentally (indirectly) restricted (because we do not have performance scores on y for those not selected on x), and (c) z is an experimental "potential" predictor incidentally restricted. (Note in the last chapter the roles of x and z were reversed when considering the x variable as a "potential" predictor, incidentally restricted.)

The Joint Correction Formulas

To bring transparency to the corrections, we again present Equations 2-35 and 2-36 from Chapter 2 that are commonly used for correcting for explicit and incidental range restriction (Lord & Novick, 1968; Sackett & Yang, 2000). Assuming direct selection on x:

$$r_{xy} = \frac{r^{*}_{xy}\left(\frac{S_x}{S^{*}_x}\right)}{\sqrt{1 - r^{*2}_{xy} + r^{*2}_{xy}\left(\frac{S^2_x}{S^{*2}_x}\right)}}. \quad (7\text{-}1)$$

On the other hand, the indirect range restriction correction formula given by (e.g., Sackett & Yang) is:

$$r_{zy} = \frac{r^*_{zy} + r^*_{xy} r^*_{xz} \left(\frac{S_x^2}{S_x^{*2}} - 1\right)}{\sqrt{\left[1 + r^{*2}_{xy}\left(\frac{S_x^2}{S_x^{*2}} - 1\right)\right]\left[1 + r^{*2}_{xz}\left(\frac{S_x^2}{S_x^{*2}} - 1\right)\right]}}. \quad (7\text{-}2)$$

Going from the restricted $r^*_{x,z,y}$ matrix to unrestricted $r_{x,z,y}$, is relatively straightforward. That is, we estimate the unrestricted population correlations by applying Equation 7-1 to solve for the unrestricted for r_{xy} and r_{xz} and Equation 7-2 to solve for the unrestricted r_{zy}.[5] If the researcher would like to go further and obtain a correlation matrix with correlation values estimated to be free of the effects of unreliability ($r_{tx,tz,ty}$), we must have the unrestricted reliabilities.

The correction sequence, as discussed, depends on the type (restricted vs. unrestricted) of the available reliability estimates (Stauffer & Mendoza, 2001). If all of the reliability estimates have been obtained in the selected sample, then we correct first for unreliability and then correct for range restriction. On the other hand, if the reliability estimates apply to the unrestricted applicant pool, we correct first for range restriction and then for unreliability (the preferred sequence). The situation requires a slightly different approach if the reliability estimates are mixed in level, some coming from the restricted sample and some from the unrestricted one, discussed below.

Before discussing the corrections in general, we address a situation involving only one correlation. Suppose that we are interested in estimating the "true" unrestricted correlation $r_{tx,ty}$. This is the unrestricted correlation between t_x and t_y. Next assume that we have the unrestricted reliability for x but the restricted reliability for y. The fact that we have a restricted estimate (an estimate from the selected sample) of the reliability of y is of no real concern if the ratio S_x^{*2} / S_x^2 is known (indicated by U in the previous chapter). We begin the process by unrestricting the reliability of y to bring the y and x reliabilities to the unrestricted level. We correct the reliability of y by modifying Equation 2-36 (7-2) following Sackett, Laczo, and Arvey (2002), as:

$$\rho_{yy} = \frac{r^*_{yy'} + r^*_{xy} r^*_{xy'} \left(\frac{S_x^2}{S_x^{*2}} - 1\right)}{\sqrt{\left[1 + r^{*2}_{xy}\left(\frac{S_x^2}{S_x^{*2}} - 1\right)\right]\left[1 + r^{*2}_{xy'}\left(\frac{S_x^2}{S_x^{*2}} - 1\right)\right]}}. \quad (7\text{-}3)$$

[5] Note that if one were to use the Expectation-Maximization Algorithm available in many computer programs, we could easily go from the restricted matrix to the unrestricted matrix without having to be concerned about direct or indirect corrections.

Lord and Novick (1968, p. 130) present the much simpler formula for obtaining an estimate of ρ_{yy}, as (changing their x variable application to y):

$$\rho_{YT}^{*2} = 1 - [(S_Y^2 / s_y^{*2})(1 - \rho_{YT}^2)]. \quad (7\text{-}4)$$

Equation 7-4 simply comes from the algebraic manipulation of the second of our classical measurement theory equalities:

$$s_e^{*2} = S_E^2 = s_y^{*2}(1 - \rho_{YT}^2), \text{ and}$$

$$s_y^{*2}(1 - \rho_{yt}^{*2}) = S_Y^2(1 - \rho_{YT}^2).$$

We note that Equation 7-4 was applied correctly in the Mayberry and Wright (1992) study reported in Chapter 4 that involved estimating the reliability of incidentally range restricted job performance measures (the criterion variables, y) for the unrestricted ASVAB population. The estimated unrestricted population y variance was accomplished using the multivariate correction for range restriction (Chapter 5 and elsewhere, applying all ASVAB tests at the population level as explicit selection variables). This correction not only yields range corrected correlations, but range corrected standard deviations and test score means. Mendoza and Munford (1987) show this added standard deviation correction for joint-correction scenario for the two-variable case.

Whatever the method, now that we have two unrestricted reliability estimates, ρ_{yy} and ρ_{xx} (ρ_{xx} assumed from the start as known), we proceed to obtain the unrestricted "true" correlation by applying the correction of r_{xy} for unreliability in both x and y:

$$r_{tx,\,ty} = \frac{r_{xy}}{\sqrt{\rho_{xx}\rho_{yy}}}, \quad (7\text{-}5)$$

which applied to Equation 7-2 to give the joint correction as:

$$r_{tx,ty} = \frac{r_{xy}^{*}(\frac{S_x}{S_x^{*}})}{\sqrt{1 - r_{xy}^{*2} + r_{xy}^{*2}(\frac{S_x^2}{S_x^{*2}})}} \frac{1}{\sqrt{\rho_{xx}\rho_{yy}}}. \quad (7\text{-}6)$$

We could, however, depending upon the research objectives, have just corrected for either the reliability of x or the reliability of y and not both. In this case, we would just modify the last portion of Equation 7-5 to accommodate only one reliability estimate.

True and Error Score Correlation: A Complication

Although there is no issue with correcting the reliability estimate of an incidentally selected *y* variable for restriction in range, there is a complicating issue for the *x* variable under direct restriction in range. If the unrestricted reliability of *x* is not available (a very unusual situation for Navy researchers, and unheard of for the ASVAB), we must proceed with caution. It has been shown that t_x and e_x when subjected to explicit selection are negatively correlated in the restricted sample (Mendoza & Mumford, 1987; Hunter, Schmidt, and Le, 2006), illustrated in Figure 7-2.

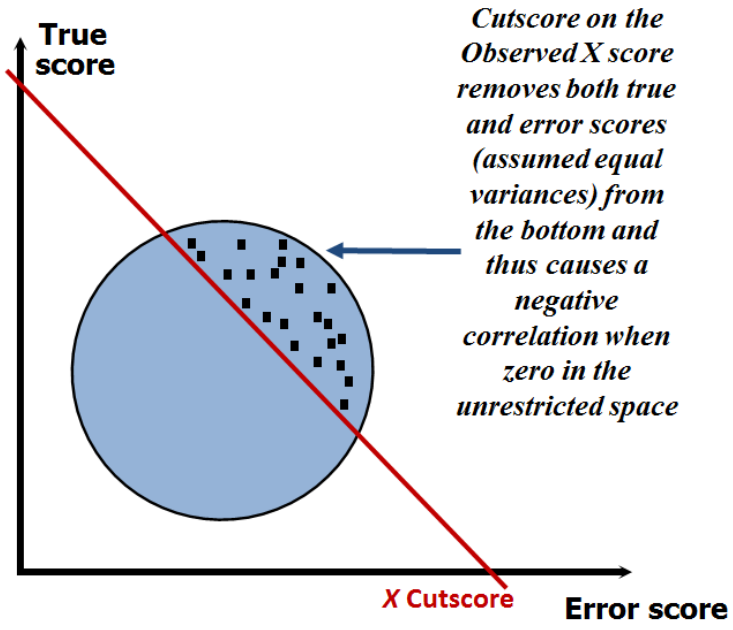


Figure 7-2. Range Restriction effects on the relation between true and error scores.

Because of the negative correlation between the error and true scores displayed in Figure 7-2 (that generally increases in magnitude with more stringent selection ratios and lower reliability), the total score variance must be expressed as

$$S_x^{*2} = S_t^{*2} + S_e^{*2} + 2\,\mathrm{cov}^*(t,e) \quad (7\text{-}7)$$

and not by

$$S_x^{*2} = S_t^{*2} + S_e^2, \quad (7\text{-}8)$$

recognizing in Equation 7-8 that error variance is not restricted when the variable involved is subject to incidental selection or when selection occurs on true scores (but unfortunately this does not apply to our *x* variable because it is subject to explicit selection). Because the covariance (and therefore the correlation) between *t* and *e* are

negative, we can see that the total *x* score variance for the restricted sample in Equation 7-7 will be *smaller* due to direct restriction in range resulting in a larger true score/total score variance ratio, thus overestimating the restricted in range reliability estimate and underestimating the unrestricted reliability estimate.

For theoretical and practical purposes, Mendoza, Stafford, and Stauffer (2000) developed a procedure for estimating the unrestricted reliability using a test-retest reliability estimate obtained in a directly restricted sample (assuming an actual re-administration of the *x* test can occur under acceptable conditions). As an independent administration of x has taken place, the error scores are not correlated between the *x* and the retest portion of *x*, say *x'*. The variance of x' reduces to the usual sum of true plus error variances; that is, Equation 7-8 holds for *x'* but not for *x* (for *x*, Equation 7-7 holds). Putting these two observations together, Mendoza, et al. (2000) showed that we can estimate the unrestricted reliability of *x* from

$$\rho_{xx} = \frac{cov*(x,x')}{S_x^{*2}}. \quad (7\text{-}9)$$

The numerator of Equation 7-9 contains the covariance between *x* and *x'* (in the selected sample), and the denominator contains the restricted variance of *x (the original administration)*, that is, the variance of the measure used for selection. Note that we are using the regression of *x'* on *x* in the restricted sample to estimate the regression of *x'* on *x* in the unrestricted sample, the unrestricted reliability. You may recall that the regression coefficient is not affected by range restriction and when parallel forms (or test-retest) have the same unrestricted variances, then their ratio equals one and thus,

$b = r\frac{S_y}{S_x}$ applied to parallel forms becomes, $b = r\frac{S_{x'}}{S_x}$, with $b = r$.

Thus, it makes sense to use the regression coefficient to estimate the unrestricted reliability. (Notice that in the unrestricted sample the correlation between *x* and *x'* is equal to the regression because the variance of *x* is equal to the variance of *x'*.) Furthermore, Fife et al. (2012) have shown in a simulation study that the approach given in Equation 7-9 of using the retest estimate in a selected sample to estimate the unrestricted reliability is unbiased.

Because ASVAB reliabilities are documented for full range groups (see Chapter 4), it should never be the case that we would need to estimate unrestricted ASVAB reliabilities from an ASVAB range-restricted school sample. However, we might need to estimate a potential predictor's full range reliability if it were used operationally (cutscore) when the only reliabilities available are from a range restricted sample (say, during the research project's time frame when both validities and reliabilities were assessed). Technically, however, because an ASVAB classification composite was the operational standard at the time of the concurrent validity study, the potential predictor is only the incidental selection variable, as is the y variable, and therefore, we are not dealing with correlated true and error scores. The point to remember is that if a variable is used for

selection and the norm (published) reliability is not available, it would be inappropriate to use the scores in the selected sample to estimate the unrestricted reliability. At the very least, we need to retest the subjects in the selected sample.

It may be a more common scenario in industry (not for the military) that an organization does not use any cognitive selection instrument in selecting personnel, and at some point decides that it should. The chosen instrument may not have published reliabilities and for convenience, the organization may only consider reliability estimation for the incumbent sample. In this case, because direct selection has occurred on this newly instated *x* variable, we have to advise that the test-retest reliability estimation will be required (or parallel forms) and not an internal consistency type of reliability due to the explicit selection effect of correlated true and error scores. On the other hand, if a cognitive measure is already in place for personnel screening (again without known reliability), there is no issue and any type of reliability estimator can be considered. At this point we refer the reader back to Chapter 4 to assess the most appropriate reliability estimation types for specific situations (we prefer stability or equivalence reliability estimators when concerned with validity coefficients).

Estimating the Restricted Reliability from the Unrestricted Estimate

We mentioned earlier that the researcher must bring the reliabilities to the same level in the correction for range restriction. For completeness, we include the rationale and formulas for estimating the restricted reliability having obtained the estimated unrestricted reliability via test-retest in the selected sample (Equation 7-9).

First, we have noticed that the unrestricted reliability is obtained from the regression of x' on x in the directly selected group

$$\rho_{xx} = \frac{cov^*(x,x')}{S^*(x)} = b_{x'x}.$$

Also, we know that the variance of the new administration in the selected group (notation now "v") is equal to the sum of the true and error variances,

$$v^*(x') = v^*(t) + v(e).$$

Notice that the variance of the errors in the new administration is not reduced. Putting these facts together, we can see that error variance can be obtained from the regression coefficient and the unrestricted variance of x,

$$v(e) = (1 - b_{x'x})V(x).$$

Thus, it follows that the restricted reliability (local) can be obtained from the variance of the new administration and the unrestricted variance as follows,

$$r^*_{xx} = \frac{v^*(x') - ((1 - b_{x'x})V(x))}{v^*(x')}.$$

Concluding Remarks

As we have seen, it is a complicated matter to estimate the unrestricted population validity coefficient free of measurement error in all variables concerned when all we have is the restricted and fallible versions. It is particularly difficult when dealing with the explicit selection variable's reliability estimate in the restricted sample. We note that the formula developments in this chapter assume that classical measurement assumptions are met, which is rarely the case, and so the reader should proceed with caution when making the joint corrections and under the specific difficult situations described in this chapter. It also should be kept in mind that as we correct for artifacts in the pursuit of unbiased estimators, precision decreases as the standard errors increase. If samples are large, this is not much of an issue, but it is an issue when samples are small.

We note that only range restriction is currently addressed in the Navy's ASVAB validation/standards studies because of the operational focus of selecting and classifying personnel. The ASVAB reliabilities are known and we take for granted that the criterion measures (training performance measured for all Navy occupations) are of high integrity. This may not always be the case so monitoring efforts will always be a requirement. The current thinking in military personnel research is that the joint corrections will be more relevant as candidate additions to the ASVAB are considered in an applied research context rather than an operational context. The methods also will be relevant if measures other than training performance become additional criterion variables in validation studies.

The next chapter deals with another complication in the evaluation of the validity coefficient that applies to the population of interest – the effects violations in the underlying range restriction assumptions have on the accuracy of the estimated population validity coefficient.

Chapter 7. References

Fife, D. A., Mendoza, J. L., & Terry, T. (2012). The assessment of reliability under range restriction: A comparison of α, ω, and test-retest reliability for dichotomous data. *Educational and Psychological Measurement, 19,* 862-888.

Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology, 91,* 594-612.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Mayberry, P. W., & Wright, W. H. (1992). *Reliability of mechanical maintenance performance measures*. (CRM 91-246/February 1992). Alexandria, VA: Center for Naval Analyses.

Mendoza, J. L., & Mumford, M. (1987). Corrections for attenuation and range restriction on the predictor. *Journal of Educational Statistics, 12,* 282-293.

Mendoza, J. L., Stafford, K. L., & Stauffer, J. M. (2000). Large-sample confidence intervals for validity and reliability coefficients. *Psychological Methods*, *5*, 356-369.

Sackett, P. R., Laczo, R. M., & Arvey, R. D. (2002). The effects of range restriction on estimates of criterion interrater reliability: Implications for validation research. *Personnel Psychology, 55,* 807-825.

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology, 85,* 112-118.

Stauffer, J. M., & Mendoza, J. L. (2001). The proper sequence for correcting correlation coefficients for range restriction and unreliability. *Psychometrika, 66,* 63-66.

Chapter 8.
Standard Errors of the Corrected Correlation
Jeff W. Johnson

Introduction

It is well known that the validity coefficient for a personnel selection test will be attenuated due to both unreliability of measures and restriction in range of test scores, as discussed in previous chapters. Little attention is typically paid, however, to the standard error of validity coefficients (correlations) corrected for range restriction. Assessing the standard error of the validity coefficient is essential for constructing confidence intervals (which vary in width with sample size, all other things being equal) and making valid inferences about the population. This chapter provides a brief review of the two major approaches reported in the literature for estimating standard errors of corrected correlations due to range restriction: (a) asymptotic sampling variance formulas and (b) bootstrapping. The chapter also provides a brief discussion of the joint correction for range restriction and unreliability.

Asymptotic Sampling Variance Formulas

Several researchers have investigated the sampling distributions of correlations corrected for range restriction and developed sampling variance formulas (e.g., Allen & Dunbar, 1990; Forsyth, 1971; Gullickson & Hopkins, 1976; Mendoza, 1993). Similarly, researchers have investigated sampling variance of correlations corrected for unreliability in one or both variables (e.g., Forsyth & Feldt, 1969; Hakstian, Schroeder, & Rogers, 1988, 1989; Mendoza, Stafford, & Stauffer, 2000; Rogers, 1976). Bobko and Rieck (1980) and Bobko (1983) derived a formula for estimating the standard error of correlations corrected for range restriction and unreliability in one variable.

Only Raju and Brand (2003) have presented a formula for estimating the standard error of correlations that have been corrected for range restriction and unreliability in both variables. This formula is useful because it is a simpler expression than previous formulas and does not require a separate formula for different definitions of reliability.

Raju and Brand's (2003) asymptotic sampling variance formula for correlations corrected for direct range restriction and unreliability in both variables is expressed as

$$\hat{V}(\hat{\rho}_{xy}) = \frac{k^2 r_{xx} r_{xy} (r_{xx} - r_{xy}^2)(r_{yy} - r_{xy}^2)}{(n-1)\hat{W}^3}, \quad (8\text{-}1)$$

where

$$\hat{W} = r_{xx} r_{xy} - r_{xy}^2 + k^2 r_{xy}^2 . \quad (8\text{-}2)$$

In these equations, r_{xy} is the observed correlation between a predictor x and a criterion y, r_{xx} is the reliability of the predictor, r_{yy} is the reliability of the criterion, and $\hat{\rho}_{xy}$ is the estimated population correlation after correction for range restriction and unreliability in both variables. In addition, k represents the ratio of the unattenuated, unrestricted standard deviation to the unattenuated, restricted standard deviation for x. That is,

$$k = \frac{S_{t_x}}{s_{t_x}} = \frac{S_x\sqrt{R_{xx}}}{s_x\sqrt{r_{xx}}}. \quad (8\text{-}3)$$

Because the reliability of x in the unrestricted sample (R_{xx}) may sometimes be unavailable, Raju, Lezotte, Fearing, and Oshima (2006) offered the following derivation of k:

$$k = \sqrt{\frac{\left(\frac{S_x}{s_x}\right)^2 + r_{xx} - 1}{r_{xx}}}. \quad (8\text{-}4)$$

It should be pointed out, however, that the type of reliability estimate used is important. Internal consistency reliability estimates of x in the restricted sample are not recommended. If one must estimate the reliability of x in the restricted sample, one should use a test-retest reliability estimate. (See Fife, Mendoza, and Terry, 2012 and the previous chapter for an explanation of the reason.)

The standard error of $\hat{\rho}_{xy}$ is $\sqrt{\hat{V}(\hat{\rho}_{xy})}$. Furthermore, Raju et al. (2006) admit that the correct estimation of k requires the assumption that t and e are not correlated, a situation not likely to hold in direct range restriction situations. If there is a second administration in the selected group, however, we can estimate k, without assuming independence between t and e in the selected group, as follows:

$$k = \frac{S_x^2\sqrt{R_{xx}}}{\sqrt{v(x') - S_x^2(1 - R_{xx})}},$$

where $v(x')$ is the variance of the new administration under the restricted space.

Raju and Brand (2003) showed that Equation 10-1 is a general formula that can be applied when there is no range restriction or when one corrects for unreliability in either x or y instead of both. For example, if a correlation has been corrected only for range restriction, r_{xx} and $r_{yy} = 1$. Equation 8-1 then reduces to

$$\hat{V}(\hat{\rho}_{xy}) = \frac{k^2(1-r_{xy}^2)^2}{(n-1)(1-r_{xy}^2+k^2r_{xy}^2)^3}. \quad (8\text{-}5)$$

This is the same equation that Bobko and Rieck's (1980) sampling variance formula reduces to when r_{yy} = 1. An equivalent but simpler formula for computing the standard error of a correlation corrected only for range restriction is

$$SE(\hat{\rho}_{xy}) = \frac{\hat{\rho}_{xy}(1-\hat{\rho}_{xy}^2)}{r_{xy}\sqrt{n-1}}. \quad (8\text{-}6)$$

It should be pointed out that regardless of how we estimate the standard error of the corrected correlation under direct range restriction (using either Bobko & Rieck ,1980 with the bootstrap, or Mendoza, 1993), if the unrestricted reliabilities are known, the standard error of the corrected correlation is given by

$$SE\left(\frac{R_{xy}}{\sqrt{R_{xx}R_{yy}}}\right) = SE(R_{xy})\frac{1}{\sqrt{R_{xx}R_{yy}}},$$

(since the reliabilities are known we can treat them as constant in the computation of the standard error.)

Assuming that sampling errors of corrected correlations are normally distributed, a *Z* test for determining whether a corrected correlation is significantly different from a hypothesized population correlation is

$$Z = \frac{\hat{\rho}-\rho}{\sqrt{\hat{V}(\hat{\rho})}}. \quad (8\text{-}7)$$

If $|Z| > |z|$, where *z* is the table's value from the unit normal distribution for a given alpha level (e.g., when α = .05, *z* = 1.96 for a two-tailed test), then $\hat{\rho}$ is considered to be significantly different from ρ. When testing whether two independent corrected correlations are significantly different from each other, the appropriate test is

$$Z = \frac{\hat{\rho}_1-\hat{\rho}_2}{\sqrt{\hat{V}(\hat{\rho}_1)+\hat{V}(\hat{\rho}_2)}}. \quad (8\text{-}8)$$

A Monte Carlo study by Raju and Brand (2003) found that the asymptotic sampling variance formula (Equation 8-1) provides accurate estimates of the sampling variance of corrected correlations. The authors also found that observed alpha levels from the formula were very similar to the nominal alpha levels, although the former consistently overestimated the latter. Also, the power rates for the formula tended to be very low, which was consistent with the power rates found in other studies of significance tests for corrected correlations (e.g., Hakstian et al., 1988).

Raju and Brand (2003) noted two implications for practitioners in the use of their formula. First, their procedure is very conservative in that a Type II error (failing to reject the null hypothesis when it is false) is much more likely to occur than a Type I error (rejecting the null hypothesis when it is not false). Second, their procedure assumes that the corrected correlations are normally distributed, which might not be the case in practice. The authors recommend continuing to develop new significance tests for corrected correlations that are based on different distributional assumptions and thus might have higher power (detecting a false null hypothesis).

Bootstrapping Approaches

Bootstrapping, mentioned in Chapter 2, is a nonparametric procedure that can be applied for estimating standard errors of any sample statistic. Using the bootstrap does not require assumptions about an underlying population distribution (i.e., normality) as does the use of parametric based procedures that use standard equations. In an ASVAB validation/standards study, the bootstrap can be applied to the school sample at hand (e.g., n= 250 records for students having complete data on both the ASVAB and final school course grade). The standard error of a bootstrapped multivariate range corrected validity coefficient is derived as the standard deviation of those validities for a large number of bootstrapped samples (e.g., 1,000) where each sample is the same size as the original sample. Each of the 1,000 samples is formed as follows. Bootstrap Sample #1 is formed from randomly drawing a case from the original sample and replacing that case (in essence, leaving the original sample intact) for the next draw until the n=250 sample is formed. The process repeats until Bootstrap Sample #1000 is formed. The standard deviation of 1,000 corrected validities is taken as the standard error (Efron, 1979).

The bootstrap approach has been shown to be appropriate for the bivariate correlation situation (Bickel & Freedman, 1981, Lunneborg, 1985). Mendoza, Hart, and Powell (1991) derived a confidence interval for a correlation corrected for range restriction ($\hat{\rho}$) based on a bootstrap procedure and investigated the accuracy and stability of the confidence interval under conditions of incomplete truncation. Incomplete truncation means that a probability mechanism is used to select cases, where those with a higher score on x have a higher probability of being selected. Incomplete truncation is a situation that is similar to what might be seen in test data that are range restricted due to indirect (incidental) selection. To study range restriction due to direct (explicit) selection, a truncation method would need to be used that produced a sample that excluded all cases falling below a cutscore.

Chan and Chan (2004) conducted a bootstrap study in which they investigated the sampling variance of the corrected correlation resulting from direct restriction in range. The experimental conditions included both (a) normal and nonnormal data and (b) different levels of ρ, selection ratio, sample size, and truncation type. The authors compared Monte Carlo and bootstrapped distribution variance with Bobko and Rieck's (1980) formula for estimating the standard error of a correlation corrected for range restriction. Recall that when correlations are corrected only for range restriction and not for unreliability, the standard error formulas given by Bobko and Rieck are equivalent to those given by Raju and Brand (2003) (see Equation 8-5). The Chan and Chan results indicated that the bootstrapped standard error is generally more accurate than Bobko and Rieck's, especially with small sample sizes. In contrast, Li, Chan, and Cui (2011) investigated indirect restriction in range showing that the bootstrap procedure produced standard errors of corrected correlations and confidence intervals that were generally more accurate across conditions including, as did Chan and Chan, combinations of sample size, selection ratio, ρ, and types of nonnormal distributions.

Concluding Remarks

A general formula developed by Raju and Brand (2003) for computing sampling variances for range corrected correlations was presented for the case of direct (explicit) selection and unreliability in both the *X* and *Y* variables. The formula is based on asymptotic sampling variance theory and therefore can be used for computing confidence intervals or testing for significance when the samples are relatively large. In the practical/operational ASVAB validation/standards setting process, it may be just as appropriate to apply the bootstrap technique. The next chapter reports on a Monte Carlo/bootstrap simulation study involving the bootstrap using the ASVAB. The study examined the accuracy of the multivariate range correction procedure for estimating unrestricted population ASVAB validity estimates under various conditions and simulation-based estimates of the standard errors.

Chapter 8. References

Allen, N. L., & Dunbar, S. B. (1990). Standard errors of correlations adjusted for incidental selection. *Applied Psychological Measurement*, *14*, 83-94.

Bickel, P., & Freedman, D. (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics, 9,* 1196-1217.

Bobko, P. (1983). An analysis of correlations corrected for attenuation and range restriction. *Journal of Applied Psychology*, *68*, 584-589.

Bobko, P., & Rieck, A. (1980). Large sample estimators for standard errors of functions of correlations coefficients. *Applied Psychological Measurement*, *4*, 385-398.

Chan, W., & Chan, D. W. L. (2004). Bootstrap standard error and confidence intervals for the correlation corrected for range restriction: A simulation study. *Psychological Methods*, *9*, 369-385.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*, 1-26.

Fife, D. A., Mendoza, J. L., & Terry, T. (2012). The assessment of reliability under range restriction: A comparison of α, ω, and test-retest reliability for dichotomous data. *Educational and Psychological Measurement, 19,* 862-888.

Forsyth, R. A. (1971). An empirical note on correlation coefficients corrected for restriction in range. *Educational and Psychological Measurement*, *31*, 115-123.

Forsyth, R. A., & Feldt, L. S. (1969). An investigation of empirical sampling distributions of correlation coefficients corrected for attenuation. *Educational and Psychological Measurement*, *29*, 61-71.

Gullickson, A., & Hopkins, K. (1976). Interval estimation of correlation coefficients corrected for restriction of range. *Educational and Psychological Measurement*, *36*, 9-25.

Hakstian, A. R., Schroeder, M. L., & Rogers, W. T. (1988). Inferential procedures for correlation coefficients corrected for attenuation. *Psychometrika*, *53*, 27-43.

Hakstian, A. R., Schroeder, M. L., & Rogers, W. T. (1989). Inferential theory for partially disattenuated correlation coefficients. *Psychometrika*, *54*, 397-407.

Li, J. C., Chan, W., & Cui, Y. (2011). Bootstrap standard error and confidence intervals for the correlations corrected for indirect range restriction. *British Journal of Mathematical and Statistical Psychology, 64*, 367-387.

Lunneborg, C. E. (1985). Estimating the correlation coefficient: The bootstrap approach. *Psychological Bulletin*, *98*, 209-215.

Mendoza, J. L. (1993). Fisher transformations for correlations corrected for selection and missing data. *Psychometrika*, *58*, 601-615.

Mendoza, J. L., Hart, D. E., & Powell, A. (1991). A bootstrap confidence interval based on a correlation corrected for range restriction. *Multivariate Behavioral Research*, *26,* 255-269.

Mendoza, J. L., Stafford, K. L., & Stauffer, J. M. (2000). Large-sample confidence intervals for validity and reliability coefficients. *Psychological Methods*, *5,* 356-369.

Raju, N. S., & Brand, P. A. (2003). Determining the significance of correlations corrected for unreliability and range restriction. *Applied Psychological Measurement*, *27*, 52-71.

Raju, N. S., Lezotte, D. V., Fearing, B. K., & Oshima, T. C. (2006). A note on correlations corrected for unreliability and range restriction. *Applied Psychological Measurement*, *30*, 145-149.

Rogers, W. T. (1976). Jackknifing disattenuated correlations. *Psychometrika*, *41*, 121-134.

Chapter 9.
A Monte Carlo/Bootstrap Study of Range Corrected Validity Accuracy

John H. Wolfe

Introduction

As we saw in the last chapter, there is error associated with a sample-based statistic used as an estimator of a population parameter. The error in the estimation of the population parameter refers to accuracy, and the range of error upon repeated trials refers to precision. This chapter describes a Monte Carlo study that examines both accuracy and precision using the ASVAB tests and the multivariate correction for range restriction (described in Chapter 5).The Monte Carlo study design involved the conditions of predictor/criterion specification, sample size, selection ratio, and distribution skew. The predictors were two ASVAB composites and the criteria were two ASVAB tests that served as surrogate criteria. The validity coefficients of the selector composites were, of course, known (referred to as "R" in some graphs, not to be confused with the multiple correlation, *R*), and so we could evaluate the accuracy and precision of the multivariate correction for range restriction under the various conditions.

Background

Several Monte Carlo studies have been conducted to examine the accuracy of the correction for range restriction in estimating the unrestricted validity coefficient. Most of the studies involved the univariate rather than the multivariate correction for range restriction. These Monte Carlo studies examined the univariate correction formula accuracy and the standard deviations of corrected validity coefficients to see what formula might apply (e.g., the standard error formula for the bivariate case). Few studies have examined the multivariate correction for range restriction, probably because the procedure is mainly applied by the military with use of all ASVAB scores that are available for all military applicants. However, it has been determined that the multivariate formulas are generally more accurate than the univariate formulas (Booth-Kewley, 1985; Held & Foley, 1994), at least for adequate sample sizes.

Accuracy of the multivariate correction for range restriction has been postulated to occur due to "…(a) inclusion of variables with adequate distributional properties, (b) the compensatory effects of regression weights, and (c) the related psychometric principle that differentially weighting a large number of correlated predictor variables has little impact on a multiple correlation. Taken from another perspective, the multivariate correction accuracy may simply be due to the fact that a regression equation with multiple relevant predictors yields a lower standard error of estimate than a regression equation with only one of the predictors" (Wolfe & Held, 2010, p. 357).

Recall that all ASVAB tests whose scores are available for the unrestricted population are entered into the multivariate correction for range restriction as explicit selection variables, even though only a subset of ASVAB tests formed into a composite serve as the explicit selection variable.

Most studies of the correction for range restriction recommend examining extreme conditions that would affect validity accuracy and the standard error, such as varying degrees of violation of the linearity and homoscedasticity assumptions that are made in performing the correction. Although Lawley (1943) relaxed the normality assumption, at least several studies have examined varying degrees of skewness because skew can cause nonlinearity (Brewer & Hill, 1969). Finally, several studies have reported that stringency in the selection ratio has a major impact on the sampling variance of the corrected validity (Mendoza & Reinhardt, 1991; Raju & Brand, 2003).

The Monte Carlo study reported here was designed to evaluate the impact of several factors on corrected validity accuracy using the multivariate formulas. The tests used in the study are those in the ASVAB. The ASVAB is the selection and classification instrument for all U.S. military services and consists of the following nine tests: General Science (GS), Arithmetic Reasoning (AR), Word Knowledge (WK), Paragraph Comprehension (PC), Auto and Shop Information (AS), Mathematics Knowledge (MK), Mechanical Comprehension (MC), Electronics Information (EI), and Assembling Objects (AO) (see Chapter 2 of the Introductory Manual for full descriptions). Two of the ASVAB tests (PC and AS) were specified as the criteria in this study so that we would know the unrestricted validity, but also because these tests reflect underlying constructs that map to many military training requirements (i.e., PC for understanding technical manuals and AS in the learning of mechanical principles and maintenance processes).

There were four main goals of this Monte Carlo study. The first goal was to determine the effects of the study-designed conditions on the accuracy of the multivariate correction for range restriction formulas in estimating known unrestricted validities. The second goal was to determine the standard error distributions that come out of the multivariate corrected validity distributions. The third goal was to determine if an ancillary measure (termed “hit rate”) could be used with some degree of confidence to identify the predictor with largest validity coefficient. The fourth goal was to determine if the bootstrap is useful in identifying the predictor with largest validity coefficient referring to the median of a bootstrapped distribution rather than the mean, or if it was more appropriate to use the point estimate from the total Monte Carlo Sample from which the bootstrap sample was derived. The hypothesis was that the median of a bootstrapped distribution reduces the influence of outliers and therefore gives a more accurate point estimate and smaller standard error.

The Monte Carlo study incorporated the following conditions: (a) selection ratio, (b) sample size, and (c) degree of skew. The validity magnitude or covariance levels were not varied systematically in this initial phase of work. These conditions could be varied along with validity difference between ASVAB composites in a subsequent phase along with predictor and criterion unreliability. Although only two ASVAB composites served as selectors, these and other ASVAB combinations served as predictors.

Monte Carlo Methods

Generation of Synthetic Populations

Segall (2004) described the Profile of American Youth (PAY97) ASVAB norming sample that was weighted to be representative of the American youth population (21,117,079 cases). Several synthetic populations were generated from the PAY97 correlation matrix by using deviate generators provided in the International Mathematical and Statistical Library (IMSL) and Cholesky factorization routines (Ackleh, Allen, Kearfott, & Seshaiyer, 2009). The procedure is described as follows:

Let $\boldsymbol{C}$ be the covariance matrix of the PAY97 sample with m variables. The Cholesky factorization of $\boldsymbol{C}$ is a lower triangular matrix $\boldsymbol{L}$ such that

$$C = LL^T \quad . \quad (9\text{-}1)$$

where the superscript T indicates the transpose of the associated matrix. Now let $\boldsymbol{Z}$ be an N by m matrix containing N cases on m uncorrelated variables, each with zero means and unit standard deviations. Let $X = ZL$. The covariance matrix of X is

$$\frac{1}{N-1} X^T X = \frac{1}{N-1} L^T Z^T ZL = L^T (\frac{1}{N-1} Z^T Z) L = L^T L = C. \quad (9\text{-}2)$$

Thus, by generating N random vectors of m variables and applying the $\boldsymbol{L}$ transformation, a random sample with the desired covariance matrix can be generated. Notice that the uncorrelated variables in $\boldsymbol{Z}$ do not have to be independently or identically distributed or have normal distributions. Using a random normal generator with N = 20 million and m = 9, a $\boldsymbol{Z}$ matrix was generated, along with an $\boldsymbol{X}$ matrix of multivariate normal cases with the same means, standard deviations, and correlations as the PAY97 population.

Karian and Dudewicz (2000) showed that a wide variety of distributions can be fitted by the four-parameter generalized lambda distribution and presented methods for generating random variables with specified skewness and kurtosis. Using these methods, $\boldsymbol{Z}$ matrices of independently and identically distributed skewed variables were generated. When the $\boldsymbol{Z}$ matrix was multiplied by the $\boldsymbol{L}$ matrix, the resulting $\boldsymbol{X}$ matrix had the same covariance matrix as the PAY97 population, but with "ASVAB" test scores that were skewed to varying degrees.

Different levels of skewness were used to generate eight $\boldsymbol{Z}$ matrices of 20 million cases of nine variables. Only the first of the $\boldsymbol{X}$ variables had the same skewness as the first $\boldsymbol{Z}$ variable. Because the $\boldsymbol{L}$ matrix is triangular, the second $\boldsymbol{X}$ variable was a weighted sum of the first two $\boldsymbol{Z}$ variables. The third $\boldsymbol{X}$ variable was the weighted sum of the first three $\boldsymbol{Z}$ variables, etc. Because of the central limit theorem, successive $\boldsymbol{X}$ variables approached closer and closer to normality, which is to say, their skewness diminished.

Table 9-1 shows the characteristics of the "skew = -1.0" population and associated distribution descriptives for the *Forward Skew* condition.

Table 9-1
Descriptive Statistics for the Forward –1.0 Skewed Population of 20+ Million Simulated Test Scores

Test	Minimum	Maximum	Mean	*SD*	Skewness	*SE*	Kurtosis	*SE*
GS	2.10	66.33	50.0018	9.99917	-1.001	0.001	0.602	0.001
AR	-14.19	71.51	50.0011	9.99495	-0.708	0.001	0.302	0.001
WK	-11.84	71.65	49.9992	10.00185	-0.715	0.001	0.316	0.001
MK	-12.87	73.28	49.9981	9.99570	-0.604	0.001	0.224	0.001
MC	-14.36	73.27	49.9999	10.00125	-0.664	0.001	0.272	0.001
EI	-9.87	75.49	49.9969	10.00584	-0.619	0.001	0.241	0.001
AO	-13.67	75.80	49.9984	10.00477	-0.575	0.001	0.212	0.001
PC	-11.19	76.91	49.9992	9.99917	-0.590	0.001	0.225	0.001
AS	-15.70	86.16	50.0008	10.00512	-0.460	0.001	0.155	0.001
VE	-8.20	75.02	49.9990	9.99798	-0.679	0.001	0.302	0.001
AFQT	-19.00	290.00	199.9976	36.37105	-0.687	0.001	0.317	0.001
GW	-7.44	137.28	100.0000	18.97468	-0.884	0.001	0.493	0.001
GM	-10.78	138.95	100.0007	18.35251	-0.829	0.001	0.441	0.001
EL	-2.44	278.38	199.9969	34.59769	-0.783	0.001	0.404	0.001
AL	9.86	493.58	349.9943	58.00127	-0.759	0.001	0.387	0.001

Notes:
AFQT = Armed Forces Qualification Test score are in standard score format.
GW for the purpose of the study = GS + WK
GM for the purpose of the study = GS + MC
EL = is the Service's Electronics composite, AR + MK + EI + GS
AL = for the purpose of the study = integer- or unit-weighted GS + AR + WK + MC + EI + AO

As Table 9-1 shows, the first "test" (GS) has the specified –1.0 skewness. In contrast, the criterion variable PC (further down the list) has a much smaller skew (-0.590), with even less skew observed for the other criterion variable, AS (-0.460). As shown, the skewness of the variables decreases almost linearly down the rows of the table. The remaining variables, VE through AL, are composites of the test scores. The ASVAB VE is a combination of 2/3WK and 1/3PC. The GW composite, formed by equal-integer weighting of the two tests, GS and WK, served as one of two selector composites in the study; GM, formed from GS and MC, served as the other.

It is worth noting that although the "tests" have the same means, standard deviations, and correlations as the PAY97 population, the ranges of scores are far greater. In particular, some of the "test" scores are negative. This extended score range occurs even in the multivariate normal simulated populations.

By reversing the order of the variables, so that AS is first and GS is last, a reversed ***L*** matrix was produced that also generated artificial test scores with the PAY97 covariances. When applied to a skewed ***Z*** matrix, the result was a set of variables with AS (a criterion variable and the last variable listed) having the highest degree of skew and GS having the least. Table 9-2 shows the results.

Table 9-2
Descriptive Statistics for Reversed –1.0 Skewed Population of 20+ Million Simulated Test Scores

Test	Minimum	Maximum	Mean	*SD*	Skewness	*SE*	Kurtosis	*SE*
GS	-12.92	79.77	49.9982	10.00120	-0.477	0.001	0.146	0.001
AR	-16.76	78.84	49.9980	9.99516	-0.487	0.001	0.151	0.001
WK	-13.80	75.14	49.9978	9.99895	-0.576	0.001	0.207	0.001
MK	-13.62	76.07	49.9955	9.99449	-0.585	0.001	0.220	0.001
MC	-13.59	76.04	49.9993	10.00226	-0.557	0.001	0.200	0.001
EI	-13.23	73.30	49.9980	10.00386	-0.635	0.001	0.251	0.001
AO	-12.53	74.09	49.9989	10.00505	-0.656	0.001	0.275	0.001
PC	-8.88	69.85	50.0003	9.99726	-0.868	0.001	0.476	0.001
AS	2.54	66.15	50.0025	10.00534	-1.000	0.001	0.601	0.001
VE	-12.94	73.35	49.9985	9.99485	-0.667	0.001	0.292	0.001
AFQT	-29.00	297.00	199.9900	36.36357	-0.609	0.001	0.257	0.001
GW	-21.11	154.27	99.9959	18.97391	-0.511	0.001	0.170	0.001
GM	-16.92	153.53	99.9975	18.35534	-0.500	0.001	0.167	0.001
EL	-26.55	302.93	199.9896	34.59795	-0.491	0.001	0.160	0.001
AL	-28.22	520.13	349.9857	57.99905	-.513	0.001	0.176	0.001

Notes:
AFQT = Armed Forces Qualification Test scores are in standard score format.
GW = GS + WK
GM = GS + MC
EL = AR + MK + EI +GS
AL = GS + AR +WK +MC +EI +AO

Monte Carlo Simulation

Conditions Studied

Monte Carlo methods were used to investigate the effects on the multivariate corrected ASVAB validity coefficients and their differences from choice of selector, choice of criterion, selection ratio, sample size, and skewness of the parent population. The alternative selectors were GW and GM. The alternative criteria were PC and AS. Five levels of selection ratio (1.0, .8, .6, .4, and .2) and eight sample sizes (50, 75, 100, 150, 225, 350, 500, and 800) were considered. The total number of combinations of these factors is 2x2x5x8 = 160, including the redundant combinations, when SR = 1.0 and GM and GW selectors are equivalent. The parent populations had 8 levels of

skewness applied in forward and reverse variable order, making 16 skewed populations, and the multivariate normal population made the 17th population (skew levels were -2.0, -1.5, -1.0, -0.5, .0.0, +0.5, +1.0, +1.5, and +2.0). In the later analyses, the normal population was duplicated and appeared once labeled as "Forward" and once again as "Reverse." Thus, the total database consisted of 160x18 = 2,880 combinations, with some duplicate or equivalent combinations. For each of these 160 combinations, 1,000 samples were drawn randomly without replacement from one of the 17 parent populations of 20 million cases.

Measures

As defined in the notes to Table 9-1 and Table 9-2, four unit-weighted composites of test scores were constructed. For each of the 1,000 samples under each condition, a covariance matrix was constructed with all the tests (predictors and criteria) and composites. Regression equations were developed using the PC and AS tests as the criteria. Multiple correlations were "fully cross-validated." That is, for each sample point, a regression estimate was constructed from the *n*-1 other points in the sample, and the correlation of these estimates with the actual criteria was computed. Finally, all of these correlations were corrected for multivariate range restriction using Lawley's (1943) procedure.

The means and standard deviations of the corrected sample validity estimates were determined and compared with their population validities. For the regression, the population validity of the sample regression equation was determined and averaged across the 1,000 samples.

Bootstrap Means and Medians

For each of the 1,000 samples, 1,000 *resamples* were drawn with replacement. The covariances within each resample were computed and the validity estimates corrected for range restriction. The means and medians of these "bootstrapped" corrected validities were computed, and their means and standard deviations were compared with those of the Monte Carlo samples' corrected validity estimates and with each other.

Percentage of Samples Correctly Identifying Best Predictor

Separate analyses were conducted for each criterion test (PC and AS). The most valid and second-most valid ASVAB predictors of these criteria (as known in the population) were compared across the 1,000 Monte Carlo samples (and separately, the bootstrap samples). The mean and standard deviation of validity differences were computed across the 1,000 samples, as was the percentage of samples that correctly showed the predictor with known highest validity in population. One analysis compared the unit-weighted composites, another compared multiple regression with the AL composite, and a third compared the ASVAB tests.

Results

Many figures containing graphs were generated for the study; however, because of their number and size, however, we present only the ones that highlight key findings.

Standard Deviations of Corrected Validities

Standard deviations were calculated for the Monte Carlo procedure's 1,000 uncorrected and corrected validities that applied to the study's various conditions. As a baseline check, for the multivariate normal case with no selection and no correction for range restriction, it was found that the observed standard deviations of the Monte Carlo generated distributions were closely approximated by the standard formula for calculating the standard error of the correlation coefficient,

$$sdev_r = (1-\rho^2)/\sqrt{N}$$

(Stuart & Ord, 1994). The standard deviation comparison results are summarized in Figure 9-1, which is a plot of the ratio of the multivariate corrected observed validity standard deviation to the formula $sdev_r$ over selection ratios (SRs) where GM (GS+MC) was the selector and AS was the criterion (PC is not listed in the legend because it also serves as a criterion). The clear trend was for the simulation-based correction using the multivariate formulas to both under-and overestimate the formula depending on the selection ratio as well as the predictor.

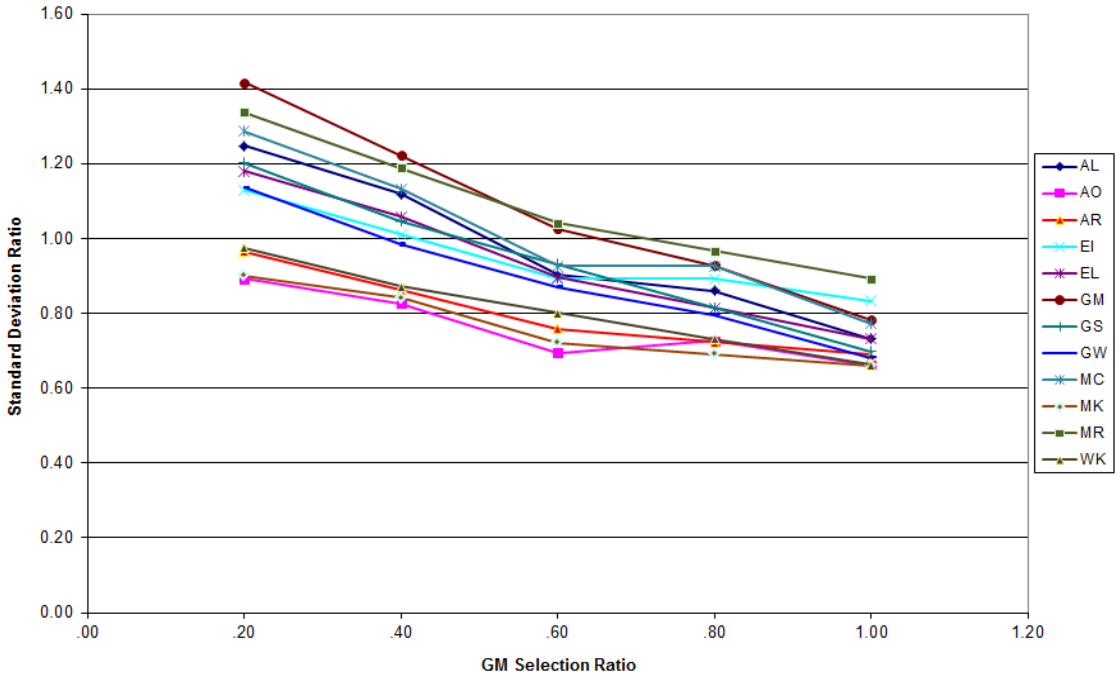


Figure 9-1. Simulated/Formula-based corrected validity standard deviation ratio with GM as the selector/predictor and AS the criterion (PAY97).

Figure 9-1 shows 12 graphed lines that apply to 12 predictors, some of which represent single tests that constitute the study's selection composites and others only incidentally correlated. (Appendix A provides the PAY97 correlation matrix for the ASVAB tests.) We first address the condition of no selection (SR = 1.0) and note that for all 12 predictors, the ratio was smaller than without selection, indicating the multivariate correction for range restriction had greater precision in estimating the population validity than the formula for $sdev_r$. We refer to "precision" as corrected validity coefficients that are within a narrow range over trials, whereas "accuracy" refers to corrected validity coefficients that are close to the known population values (Wolfe & Held, 2010).

The topmost line in Figure 9-1 at SR = 1.0 applies to the full multiple regression equation (MR) involving all ASVAB predictors, which intuitively would not be expected to follow the formula for a correlation coefficient's sampling distribution standard deviation. Relative to the other lines across selection ratios, both MR and the explicit selection composite GM (GS + MC) were higher; however, all lines trended upward over increasingly stringent SRs. At a SR = .60, both MR and GM had an approximate 1.0 ratio, indicating equality of standard deviation methods.

At a more stringent SR = .40 in Figure 9-1, most ratios exceeded 1.0, indicating a switch in the precision of the observed and formula-based standard error estimates (the formula-based method having greater precision). Also noted was the fanning out of the ratios at the most stringent SR = .20. The increase in the ratio index and the increased spread of the values among predictors may simply be due to the nature of isolated upper tail segments of a bivariate normal distribution: They are less representative of the total distribution, even with large samples (in this case, n = 800, a large enough sample size to expect stable results).

Chapter 11 provides a discussion of the potential for the multivariate correction for range restriction to produce negative corrected validity estimates in small samples when the sign is positive in the population. The phenomenon discussed in Chapter 11 applied to the current study for the smallest small sample size of 50 and a stringent SR = .20. Table 9-3 gives the unrestricted validity coefficients for the two focal ASVAB predictors in the study and the two criteria.

Table 9-3
Population Validity Coefficients

	Criterion	
Predictor	**AS**	**PC**
GM	.65	.71
GW	.50	.78

Notes: GM = GS + MC; GW = GS + WK. Test names are given on the second page of this chapter and test descriptions in Chapter 2 of the Introductory Manual.

Next, we attempted to account for the differences in the validity standard deviation ratios among predictors by relating them to the squared population validity coefficients. Figure 9-2 depicts a simplified situation where there is no selection and where the sample size is the largest in the study (n = 800).

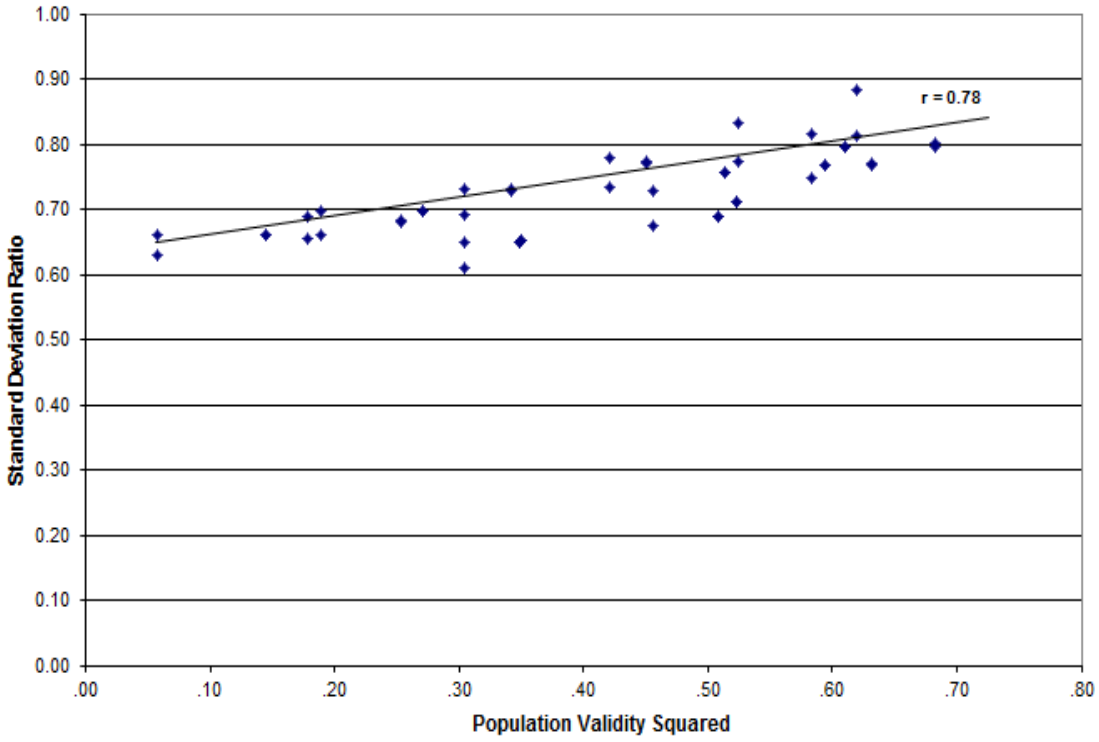
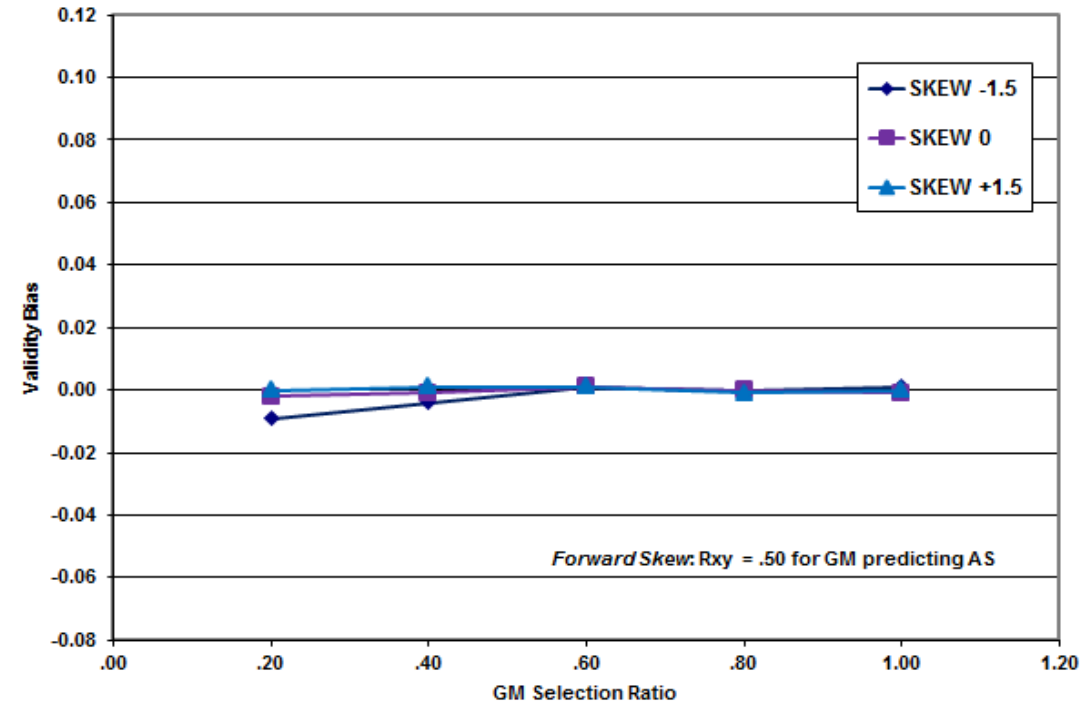


Figure 9-2. Relation between standard deviation ratios and squared population validity coefficients among predictors (no selection; n = 800).

Figure 9-2 depicts a correlation of .78 between the ratio index and the population's squared validity coefficients across the various predictors with varying known levels of predictive validity. It appears that smaller population validity coefficients are associated with smaller standard errors from multivariate range corrections compared to the standard error formula, but it is not apparent what might be the reason.

Effect of Predictor and Criterion Skew on Range-Restriction Corrections

Recall that with the *Forward Skew* condition in the population, the predictors have the greatest skew, whereas the criteria have the least skew. In the *Reverse Skew* condition, the situation is reversed. Figure 9-3 allows us to examine the effects of skew sign and magnitude on validity accuracy (average bias in the simulations) for both conditions for the study's selection ratios (the GM composite as the selection variable), the largest sample size (n = 800), and three of the study's skew values (-1.5, 0, +1.5).



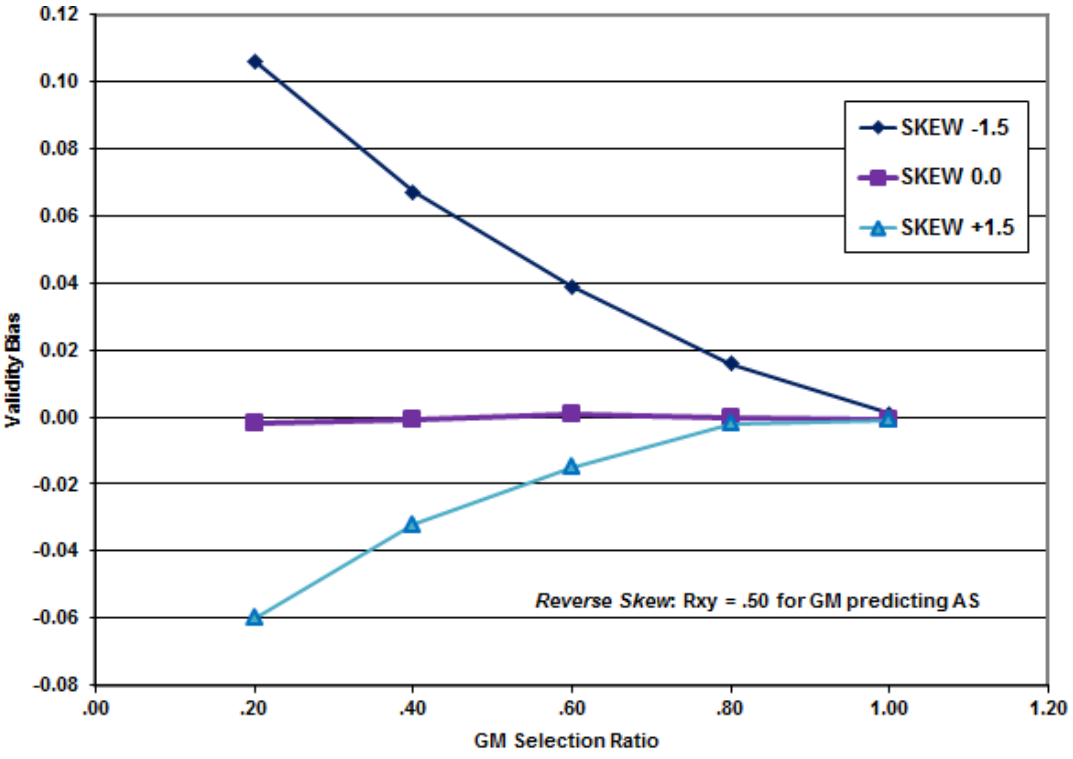


Figure 9-3. Sample validity bias across selection ratios with GM selection and GM predicting AS (*Forward* and *Reverse* Skew, *n* = 800).

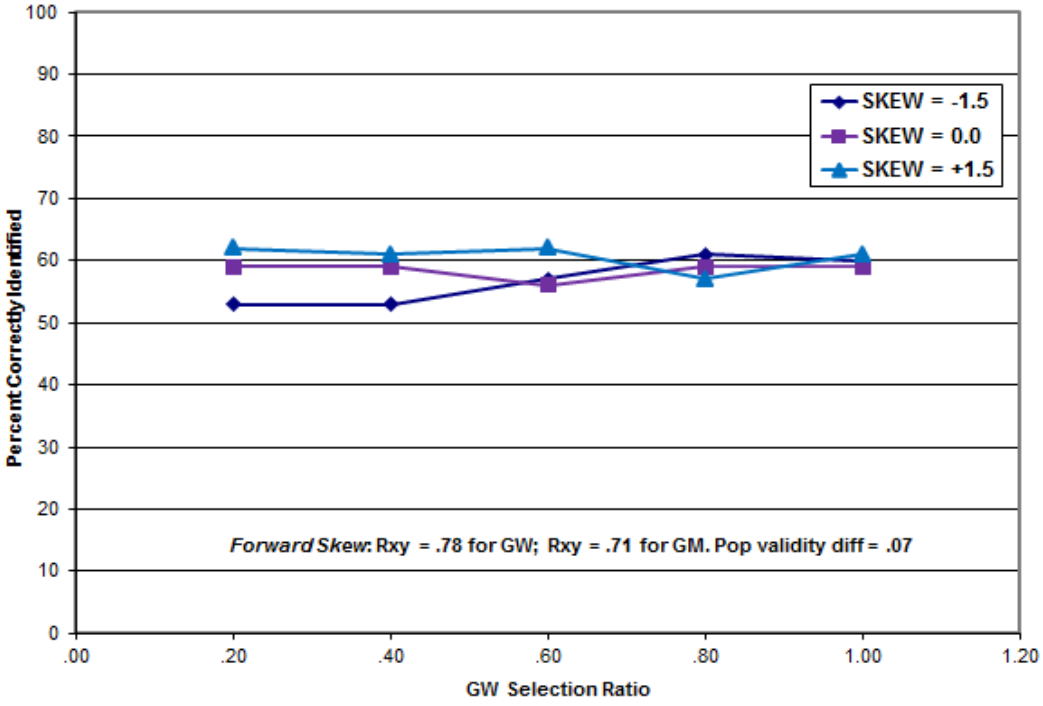
The top graph in Figure 9-3 applies to the Forward Skew condition (larger skew on the predictors and less on the criterion) and shows relatively no bias in the multivariate range correction at zero skew and over the span of increasingly positive skew values. There is, however, negative bias for negative skew, with more bias with increasingly stringent selection ratios. In contrast, the bottom graph in Figure 9-3 applies to the Reverse Skew condition (larger skew on the criterion and less on the predictors). Figure 9-3 shows a systematic increase in bias at increasingly stringent selection ratios with large positive bias associated with the -1.5 negative skew condition and slightly lower bias, but negative, for the positive skew condition.

Next, we addressed whether it is possible to identify the "best" composite in a small sample under skewed conditions. Identifying the predictor with the largest validity coefficient is always a concern in small samples due to large sampling error, but we added the data condition of skewness to gather more insight into how predictor and criterion score distributions affect the already potentially unstable validity results in a small sample. It was important to consider which predictor/criterion pair to use in demonstrating the best composite identification percentage, because the magnitude of the population validity difference depended on which criterion was used (AS or PC). We chose the PC criterion for both GM and GW to predict because the population validity differences were smaller than when AS was the criterion (.78 - .71 = .07 for the GW – GM validity difference when PC was the criterion compared to .65 - .50 = .15 when AS was the criterion).

As many readers of this document may know, with use of the current ASVAB, typically the range of incremental validity provided by the optimal composite over one in operational use is not much more than .02 to .05 (assuming final school grade in training is the criterion). However, the fact that many of the current military ASVAB composites are highly correlated at this time does not mean that they will be in perpetuity.[6] Further, identifying the best composite (speaking only in terms of validity coefficient magnitude in this chapter) would seem to be as important if not more so given the sample size is very small in a particular ASVAB validation/standards study and there is not much confidence that the magnitude of multivariate range corrected validity coefficients actually reflect the population values.

Figure 9-4 shows the accuracy of identifying the "best" composite predicting PC (with GW as the selection variable this time) as the percentage of 1,000 Monte Carlo samples that GW was identified as having a larger validity coefficient (.78) than GM (.71) resulting from the multivariate range correction. As with Figure 9-3, Figure 9-4 shows both the *Forward* and *Reverse* skew conditions as separate graphs. What differs in Figure 9-4 is that that the *y*-axis is "Percent Correctly Identified" rather than "Validity Bias" (in Figure 9-3). Also, the sample size now is the smallest $n = 50$ compared to the study's largest $n = 800$, the selector is GW rather than GM, and the criterion is PC rather than AS.

[6] As recommended by the ASVAB review panel (Drasgow, Embretson, Kyllonen, & Schmitt, 2006), several potentially new ASVAB classification tests are being validated for possible inclusion in a future ASVAB.



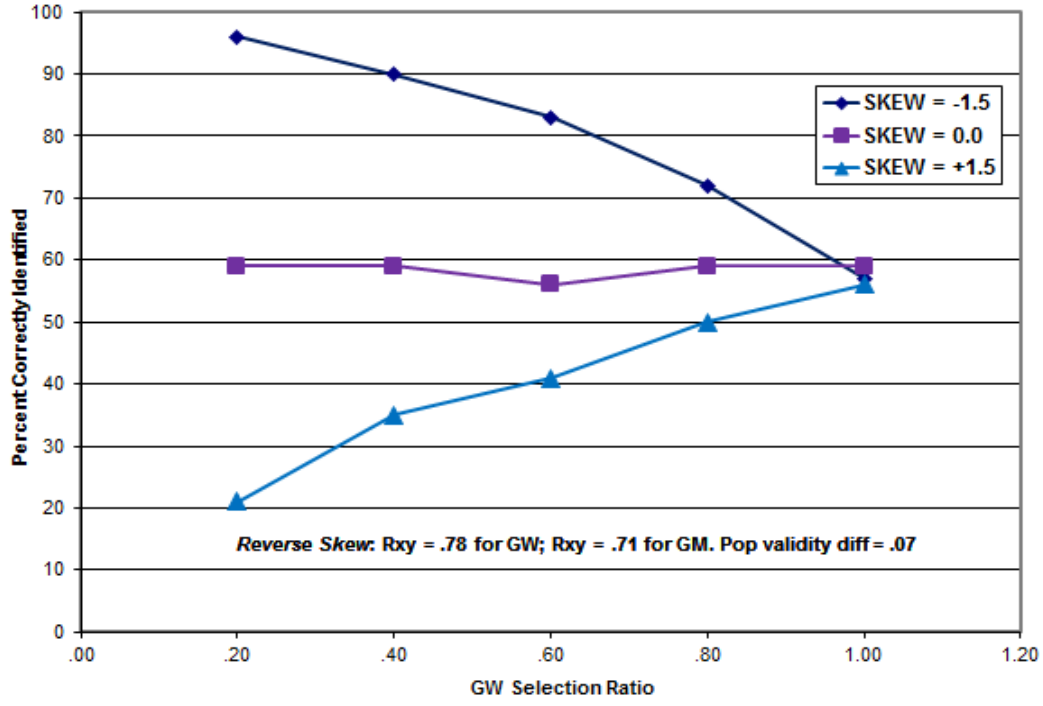


Figure 9-4. Best composite identified with GW selection comparing GW and GM predicting PC (*Forward* and *Reverse* skew, n = 50).

As seen in Figure 9-4, the forms of the two *Forward* and *Reverse* skew conditions mirror the forms in Figure 9-3. That is, the impact of large skew on the set of predictors treated as explicit selectors in the multivariate range correction is largely inconsequential to both validity bias over increasingly stringent selection ratios but also in identifying the best predictor over selection ratios, but *is* consequential when large skew is placed on the criterion. When the load of skew is reversed in this study (criterion getting the most), there are extreme results with both indices (validity bias and identifying the best composite) regardless of sample size (Figure 9-3 with n = 800, Figure 9-4 with n = 50). We note that with the n = 800 sample, GW with validity of .78 was identified 100% of the time over GM with .71 validity, showing that a sample size as small as 50 has drawbacks, at least in this simulation study without the benefit of having a more adequate real life criterion variable (e.g., the Navy's final school grade for measuring training performance). We also note again that the validity difference of .07 to detect was not set, but a feature of the ASVAB tests chosen to serve as surrogate criterion variables and the ASVAB composites chosen to serve as the selectors/ predictors.

By now, it should be clear that the effect of skewness on the accuracy of range-restriction corrections is not a simple one. It involves complex interactions with selection ratio, sample size, and the particular selectors and criteria used. The magnitude and direction of the effects seem to depend on the particular combinations of selectors and criteria. Further, from sampling theory, sample size will play a part even in the best of conditions.

We further investigated the role of sample size in the ability to at least detect the "best" composite assuming our data conditions were perfect. Table 9-4 shows the "hit" rate over all of the study's selection ratios for three pairs of the ASVAB predictors and PC as the criterion: (a) GW (GS + WK) compared to EL (AR + MK+ EI+ GS, (b) MR (optimal regression-weighted GS + AR + WK + MK + EI + MC + AO) compared to the same tests unit weighted, and (c) the single ASVAB tests, WK compared to AR.

Table 9-4 shows several interesting results for data that do not violate the multivariate range correction assumptions (i.e., the data are multivariate normal data). First, the predictor with the largest population validity coefficient is, not surprisingly, the multiple regression "best fit" equation, MR (table notes give population multiple correlations) with R_{xy} = .826. The comparison validity applies to the unit weighted version of the equation (AL) with R_{xy} = .795 for a population validity difference of .031. Detecting MR as the best predictor at a 94% accuracy range (a form of power) requires a sample size of 150. In contrast, the single ASVAB test pair (WK – AR) requires n= 350. The lower sample size required for the MR/AL validity difference detection is consistent with the larger MR and AL validity in predicting PC as compared to WK and AR, and with the "precision" of MR and AL in terms of lower standard errors in their multivariate corrected validity estimates.

Table 9-4
"Hit Rate" with and without Selection for Best Predictor across Sample Sizes with PC as the Criterion (1,000 Monte Carlo Replications)

Sample Size	Predictor Pairs		
	GW – EL %>0 Diff	MR – AL (%>0 Diff)	WK – AR (%>0 Diff)
Selection Ratio = 1.00, no selection			
50	59	35	72
75	62	54	77
100	63	72	79
150	70	94	85
225	69	99	90
350	74	100	94
500	79	100	96
800	84	100	99
Selection Ratio = .20 based on GM selection			
50	59	44	69
75	60	59	74
100	61	74	78
150	66	91	83
225	68	99	88
350	74	100	93
500	78	100	96
800	83	100	98

Notes:
GW = GS + WK with R_{xy} = .781; EL = AR + MK + EI + GS with R_{xy} = .771 for a .010 validity diff.
MR = regression weighted GS + AR + WK + MK + EI + MC + AO with R_{xy} = .826;
AL = unit weighted tests with R_{xy} = .795 for a .031 validity diff. WK R_{xy} = .764; AR R_{xy} = .723 for a .041 validity diff.

The second point of interest in Table 9-4 is that even with a stringent selection ratio (recognizing that multivariate normal without skew was used in the simulations), the results across the three predictor pairs are not too dissimilar. For example, the 150 and 350 sample size requirement for at least a 90% hit rate is the same for both the unrestricted population and the SR = .20 selected samples.

The last point of interest we discuss in Table 9-4 is the relatively low hit rate for the GW/EL validity comparison where the population difference is a small .010. Although this is a small validity increment, it is typically observed in ASVAB validation/standards studies and could be a deciding factor as to which composite to recommend. In this small validity difference case at relatively large validity coefficient magnitudes, a sample of n = 500 was required to approach the .80 hit rate in identifying the GW as the best

predictor. In ASVAB validation/standards studies, a hit rate as small as two-thirds could be considered adequate, and we see that n = 150 gives us this degree of confidence (i.e., 66% hit rate even when selection is stringent). Finally, we remind ourselves that the comparisons in Table 9-4 involve only two predictors and that the military tends to evaluate many more in a particular study. Chance factors will diminish our confidence the more composites we compare, although perhaps not to any substantial degree if we are not so concerned with the magnitude of the validity difference but instead with only which composite is "best." [7] Nevertheless, the greater the number of comparisons and the smaller the population validity difference between composites as well as the magnitude of the validity coefficients (larger being better), the greater the likelihood that a wrong composite will be identified as best, all other things being equal.

Multiple Regression vs. Unit-Weighted Composites

Unit-weighted composites are used in practice because it is believed that they will be more stable and generalizable than multiple regression derived weights, especially in small samples. In this study's simulations, multiple regression always had a larger fully cross-validated validity coefficient than any unit-weighted composite, or single test, when the sample size was 100 or greater, lending support for regression weighted composites using all of the ASVAB tests over unit-weighted composites with some small number of ASVAB tests. For sample sizes of 50, however, regression was always superior for predicting AS but was second-best for predicting PC. For sample sizes of 75, regression was usually (but not always) superior for predicting PC. These mixed findings do not support either regression-based weights or unit weights with small samples, but do with much larger samples.

Also, in addition to the corrected cross-validity estimates, it was possible to compute the population validity values (using full range population data) using the sample regression equations and unit-weighted composites. These population values were often lower than the multivariate range corrected validity estimates derived from the sample. The result that multiple regression in the population sometimes was not as good for predicting PC as one of the composites or tests in the sample, regardless of how large the sample size was, occurred mostly when the selection was stringent and the skewness was large, complicating the matter.

Because the population validity is a better index of the actual predictive value of a predictor than corrected validity coefficients, which are only estimates, these findings suggest that unit weights may be superior to regression weights when skewness and selection are extreme. Unfortunately, the population validity is known only in simulations such as those performed in this study, never in practice. We encourage the reader to explore the literature on regression versus unit weights recognizing that the Army has taken the position that full ASVAB regression weighted equations have the most utility for their enlisted classification systems.

[7] Dr. Daniel O. Segall reprogrammed a Fortran version of the multivariate range correction Fortran program to output a square matrix of the percentage of times (out of 1,000 bootstraps applied to the study sample) that one ASVAB composite had a larger corrected validity coefficient than any other included in the study.

Bootstrap Means vs. Medians

One of the purposes of this study was to determine if bootstrapping could help identify the best of several alternate predictors by comparing the median versus the mean derived in the bootstrap samples. The idea is that by virtue of selecting the bootstrap median instead of mean as the central tendency statistic, outlier values would have a diminished effect on the corrected validity coefficient. Table 9-5 shows the descriptive statistics and the bias that applies to these two central tendency statistics (i.e., population minus corrected validity estimates) resulting from the 2,880 simulations that incorporated the study conditions.

Table 9-5
Descriptive Statistics for Bias Associated with Sample Corrected Validities across 2,880 Simulation Conditions

	N	Minimum	Maximum	Mean	SD	Skewness	Kurtosis
SAMPLE	2,880	-0.233	0.132	-0.00351	0.029312	-0.308	8.263
MEAN	2,880	-0.326	0.130	-0.00874	0.035965	-2.053	13.381
MEDIAN	2,880	-0.238	0.132	-0.00295	0.029370	-0.574	9.080

In Table 9-5, SAMPLE refers to the mean bias generated from the Monte Carlo sample, whereas MEAN and MEDIAN refer to bias associated with the mean and median corrected validity coefficients generated from the bootstrap of each Monte Carlo sample. The Mean column in Table 9-5 could be used to evaluate correction bias.

Table 9-5 shows that, for the Monte Carlo results (SAMPLE), the mean bias of estimating the population validity coefficients resulting from the range restriction correction is -.00351, which is close to the value of zero we would expect if there was no bias. The mean bias across all of the individual bootstrapped Monte Carlo samples based on the mean of each bootstrapped distribution (MEAN) is -.00874, slightly larger in magnitude than that observed from the parent Monte Carlo (-.00351). In contrast, the mean bias based on the median of each bootstrapped distribution (MEDIAN) is smaller (but possibly trivially so) at -.00294. The bootstrap MEAN value appears more biased than the bootstrapped MEDIAN value when comparisons are made to the Monte Carlo mean (SAMPLE), although the practical difference (3rd decimal place) may (or may not) be considered trivial.

Consistent with the range-corrected validity results supporting the bootstrap MEDIAN to be used as the central tendency index to bootstrap, Table 9-5 also shows that the bootstrap MEDIAN's standard deviation (*SD* = .029370) is comparable to the Monte Carlo (SAMPLE) *SD* (.029312), whereas the bootstrap Mean's *SD* (MEAN) is larger (.035965). The bootstrap MEAN's distribution also has larger Skewness and Kurtosis than the two counterpart distributions, indicating outlier influence.

Table 9-6 shows the intercorrelations between the SAMPLE, MEAN, and MEDIAN corrected estimates.

Table 9-6
Correlations Between Monte Carlo Sample Validity Means, Bootstrap Means, and Bootstrap Medians across 2,880 Simulation Conditions

	SAMPLE	MEAN	MEDIAN
SAMPLE	1.000		
MEAN	.956	1.000	
MEDIAN	.998	.968	1.000

Table 9-6 shows that the correlation between corrected validity estimates is largest when derived for the Monte Carlo (SAMPLE) and bootstrap procedure that uses the median (MEDIAN) rather than the mean (MEAN). Based on the results presented in Tables 9-5 and 9-6, the median value of range-corrected validity estimates, not the mean, could be considered the appropriate bootstrap statistic when establishing the standard error of the bootstrap distribution and the construction of confidence intervals

Concluding Remarks

We have reported only some of the findings from the complete study described in this chapter, but they are of considerable interest for ASVAB validation/standards researchers. First, as noted in Chapter 2, many factors affect the correlation coefficient, and the ones of most concern are the integrity of the criterion and the distributional properties of the variables. We know a great deal about these features for the ASVAB but not for the performance measure that is used to validate the ASVAB. One of the goals of the study was to determine if the experimental conditions affected the accuracy of the multivariate correction for range restriction, and some conditions did more than others. Second, in general, a disproportionate amount of skew on the criterion relative to the predictors led to both overestimates and underestimates of the validity coefficient, all other things equal.

We also saw that the Monte Carlo-generated sampling errors of the corrected validity coefficients were both larger and smaller than the standard formula that applies to bivariate correlation, depending upon the stringency of selection under multivariate normal conditions. Fourth, across all of the study conditions, the bootstrap median corrected validity coefficient provided a very slight improvement in population validity estimates and also in reducing the standard deviation of the estimates, presumably from reducing the effects of outlier values.

We also explored sample size effects without skew effects and saw that if the objective was to identify the "best" composite out of two, a sample size of 500 would be required if we had to recommend with a high degree of certainty, say 80% (or a power of .80) that we had made the correct decision. Of course, the validity levels we reported were large and so detecting the best composite for low magnitudes validity coefficients would require even larger sample sizes (addressed in the next chapter).

Finally, we might have a better ASVAB validation situation when using the military's training grades as the criterion than is depicted by this simulation study for multivariate normality (without skew). That is, at least for the Navy, final school grade in training reflects better differentiation in individuals' performance than might be expected at stringent selection ratios imposed on a bivariate or multivariate normal distribution. But even though the military does not typically encounter the extreme conditions that were constructed in this simulation study, practitioners should be aware of them when conducting ASVAB validation/standards studies. Understanding the interactions of the relevant selection factors should be an important research goal.

The next chapter provides a section on the power of establishing a validity magnitude effect and the issues involved in comparing validity coefficient differences in personnel selection situations. The chapter also provides a brief discussion about regression methods, with the following chapter considering a variable's suppressor effects that enhance a full least squares regression equation's predictive optimization, as was observed in the study results reported in this chapter.

Chapter 9. References

Ackleh, A. S., Allen, E. J., Kearfott, R. B., & Seshaiyer, P. (2009). *Classical and modern numerical analysis*. Boca Raton, FL: Chapman & Hall/CRC Press.

Booth-Kewley, S. (1985). *An empirical comparison of the accuracy of univariate and multivariate corrections for range restriction* (NPRDC-TR-85-19). San Diego: Navy Personnel Research and Development Center.

Brewer, J. K., & Hills, J. R. (1969). Univariate selection: The effects of size of correlation, degree of skew, and degree of restriction. *Psychometrika, 34,* 347-391.

Drasgow, F., Embretson, S. E., Kyllonen, P. C., & Schmitt, N. (2006). *Technical review of the Armed Services Vocational Aptitude Battery (ASVAB)* (FR-06-25). Alexandria, VA: Human Resources Research Organization.

Held, J. D., & Foley, P. P. (1994). Explanations for accuracy of the general multivariate formulas for correcting for range restriction. *Applied Psychological Measurement, 18*, 355-367.

Karian, Z. A., & Dudewicz, E. J. (2000). *Fitting statistical distributions*. NY: CRC Press.

Lawley, D. (1943). A note on Karl Pearson's selection formula. *Royal Society of Edinburgh, Proceedings, Section A, 62,* 28-30.

Mendoza, J. L., & Reinhardt, R. N. (1991). Validity generalization procedures using sample-based estimates: A comparison of six procedures. *Psychological Bulletin, 110,* 596-610.

Raju, N. S., & Brand, P. A. (2003). Determining the significance of correlations corrected for unreliability and range restriction. *Applied Psychological Measurement, 27,* 52-71.

Segall, D. O. (2004). *Development and evaluation of the 1997 ASVAB score scale* (Technical Report No. 2004-002). Seaside, CA: Defense Manpower Data Center.

Stuart, A., & Ord, J. (1994). *Kendall's advanced theory of statistics: Distribution theory* (6th ed., Vol. 1). NY: Halsted Press.

Wolfe, J. H., & Held, J. D. (2010). Standard errors of multivariate range-corrected validities. *Military Psychology, 22*, 356-366.

Chapter 10.
Assumption Violation Effects on Range Correction Accuracy
Jeff W. Johnson

Introduction

As if there were not already enough statistical corrections we need to be concerned about in estimating our population ASVAB validation coefficient (i.e., corrections for both restriction in range and unreliability - measurement error), we now consider factors that impact the accuracy of these corrections. The joint corrections worked through in Chapter 6 do not consider sampling error or that the assumptions for correcting for range restriction may not have met (i.e., linearity, homoscedasticity, and explicit selection in the simple bivariate case of *X* and *Y* variables). This chapter focuses solely on the effects of assumption violations on the accuracy of the range corrected validity coefficient.

Background

To review briefly, if a group has been selected solely on the basis of their scores on some variable *X*, then this is known as explicit selection on *X*, and *X* is referred to as the explicit selection variable (Lord & Novick, 1968). When the variance of *x* in the selected group is smaller than the variance of *X* in the population, as is usually the case, the correlation between *x* and a criterion (*y*) in the selected group (r_{xy}) will underestimate the population correlation (R_{xy}). Pearson (1903) developed a correction formula for estimating R_{xy} given the selected group's correlation and the ratio of the variance of the predictor in the selected and total groups (s_x^2/S_x^2) (see Chapter 2's Table 2-4).

Another way that sample variability may be restricted is in the case of incidental selection (Lord & Novick, 1968). Suppose *Z* is a proposed predictor of *Y*. The range of both *Y* and *Z* is restricted to the extent to which they are both correlated with the explicit selection variable, *X*. Both *Y* and *Z* are then termed incidental selection variables. Pearson (1903) also developed a formula for the estimation of the population correlation between *Y* and *Z,* presented in Chapter 5 in the section about the trivariate case (the incidental/indirect restriction in range case).

Sackett and Yang (2000) developed a classification scheme for different range-restriction scenarios and we have recommended a read on their work in prior chapters. These scenarios were based on various combinations of the following facets: (a) variable(s) on which selection occurs, (b) whether unrestricted variances for relevant variables known, and (c) whether a possible third variable is measured or unmeasured.

The interested reader is advised to consult this article if the selection situation is more complex than the explicit or incidental situations described here.[8]

Review of Range Restriction Correction Assumptions

Correction formulas for explicit and incidental range restriction are based on several assumptions that have been previously discussed but are reiterated here: (a) there is linearity of regression of *Y* on *X*, (b) the conditional variance of *Y* given *X* is constant at all values of *X*, and (c) selection is based solely on *X*. The "a" and "b" are considered distributional assumptions and "c" a selection assumption. An additional assumption in the three-variable case is that the covariance of *Y* and *Z* given *X* does not depend on *X* (Lord & Novick, 1968). The three-variable case is the simple case of the general multivariate case, and many more X and many more Y variables can apply in one multivariate range restriction correction. Lawley (1943) relaxed the normal distribution assumption in the multivariate case, and this makes intuitive sense because we know that many inferential statistics are robust to violations of normality and that linearity can exist between variables that are not exactly normally distributed.

Several authors (e.g., Ghiselli, 1966; Guion, 1965; Linn, 1968; Lord & Novick, 1968) have noted that test score data often fail to satisfy the assumptions of linearity and homoscedasticity. Lee and Foley (1986) showed empirically that the slope of the regression line and the dispersion of *Y* on *X* are often not constant throughout the range of test scores. Therefore, it is reasonable to question the accuracy of the corrections when these assumptions are violated.

Studies of Assumption Violations

Greener and Osburn (1980) simulated test score data to examine the effect of assumption violations on the accuracy of the bivariate correction formula for explicit selection. They studied three general types of distributions: (a) sigmoid, (b) football, and (c) fan. Sigmoid distributions violate the assumption of linearity because of flattening in both tails of the bivariate distribution. Lee and Foley (1986) found that the relation between the Armed Forces Qualification Test (AFQT) and the ASVAB Mathematics Knowledge (MK) approximated a sigmoid; however, we remember that the AFQT, while sometimes appearing normally distributed in a recruit population because high aptitude youth tend to seek college options and low aptitude youth do not qualify for the military service), is scored on the percentile metric so the investigation bivariate normality was not totally appropriate. Football-shaped distributions violate the homoscedasticity assumption because the conditional variance of *Y* given *X* is at its maximum in the center and decreases in the tails of the distribution. Fan-shaped distributions also violate the assumption of homoscedasticity in that the conditional variance of *Y* given *X* increases systematically from one tail to the other.

---

[8] Note that there is a typo in Sackett and Yang's (2000) Equation 7 presenting the corrected population variance-covariance matrix in the case of multivariate correction. The value in the upper right quadrant of the matrix should be $V_{p,p}V_{p,p}^{*-1}V_{p,n-p}^{*}$ (P. R. Sackett, personal communication, March, 2000).

Greener and Osburn (1980) found in their investigations that, with a sigmoid distribution, the corrected correlations increasingly underestimated the population correlations as the degree of truncation of the distribution increased (i.e., the selection ratio became more stringent). In all cases, however, the corrected correlation was a more accurate estimate of the population correlation than was the uncorrected correlation. The trend in the fan distribution was for the corrected correlation to gradually underestimate the population correlation to a greater extent as a function of the degree of truncation and the size of the population correlation. The corrected correlation was superior to the uncorrected correlation in all cases.

With the football distribution, the correction overestimated the population correlation in samples that were truncated 50% or more. There was also a tendency for corrected correlations to overestimate the population correlation as a function of its size. The corrected correlation was not as accurate as the uncorrected correlation in highly restricted samples. The general conclusion was that for moderate degrees of restriction (i.e., selection ratios of .60 or higher), the corrected correlation is a reasonably good estimate of the population correlation and is much better than the uncorrected correlation. For more restrictive selection ratios, the corrections become progressively worse in estimating the population correlation, and unacceptable overestimates result from the football distribution.

Although Lawley (1943) relaxed the normality assumption in the multivariate correction for range restriction, Brewer and Hills (1969) showed that skewed distributions can affect linearity, and therefore the accuracy of the estimation of the population correlation.

Holmes (1990) developed a mathematical framework to investigate the effects of violations of the assumptions of linearity and homoscedasticity. Simple expressions were derived algebraically for both the selected group and corrected correlations in terms of the population correlation in the sigmoid, fan, and football situations. By plugging in preset parameter values, Holmes was able to determine what selected group magnitude of correlation r_{xy} could be expected for different levels of unrestricted validities R_{xy} values and different degrees of selection (akin to Table 2-4). The results were very similar to the results of Greener and Osburn (1980).

Gross and Fleischman (1983) studied distributions that simultaneously violated the assumptions of linearity, homoscedasticity, and selection only on *X*. They found that the correction was not robust with respect to these simultaneous violations, but it was more accurate than the uncorrected correlation much of the time. Gross and Fleischman (1987) found that the correction formula performed poorly for certain nonlinear regression forms and recommended that it should not be used unless the population correlation was thought to be large and sample sizes were large.

Johnson and Sager (1991) conducted the only study that has investigated the effects of assumption violations on correlations corrected for range restriction due to incidental selection. The authors generated simulated test score data that approximated sigmoid, fan, and football distributions, and observed the effects of different levels of selection and correlation between the explicit selection variable and each of two incidental

selection variables. Similar to the explicit selection case (e.g., Greener & Osburn, 1980), when the assumption of linearity was violated by flattening in both tails of the *XY* distribution, it appeared to be safe to correct for range restriction in all cases. The corrected correlation was at least as good an estimate as the uncorrected correlation except in the most extreme conditions.

Things were not so simple, however, when the assumption of homoscedasticity was violated. With the football-shaped distribution, the corrected correlation usually more closely estimated the population correlation than did the uncorrected correlation. The corrected correlation, however, was also much more likely to be an overestimate. Very large overcorrections usually were found only at low selection ratios.

The most troubling distribution was the fan distribution. Both the corrected and uncorrected correlations had a tendency to overestimate the population correlation when $R_{XZ} < R_{YZ}$. This phenomenon, however, was more pronounced for the corrected correlation. The uncorrected correlation in this case more closely estimated the population correlation. Correcting for range restriction, however, improved the estimate dramatically when $R_{XZ} > R_{YZ}$.

Offsetting Violations

As depicted in the previous section, the correction for range restriction's accuracy depends on adherence to the underlying assumptions for performing them, the severity of the violations, the degree of restriction in range, and sample size. Further, the influence of assumption violations on corrections for range restriction is actually more complex than is evident when violations are examined separately. For example, the linearity violation may result in an overestimate of the population (unrestricted) validity coefficient when the heteroscedasticity assumption holds, but if each assumption is violated in certain ways, they could offset each other and result in an accurate validity estimate. The offsetting linearity and heteroscedasticity violations can be evaluated in the reduced form of the unrestricted validity coefficient:

$$R_{XY} = \frac{S_{XY}}{S_X S_Y}, \quad (10\text{-}1)$$

where R_{xy} represents the corrected (estimated) population validity derived from the known sample regression coefficient b, the known population standard deviation S_x, and the derived (estimated) standard deviation. In a further reduced form of Equation 10-1

$$R_{XY} = \frac{bSx}{S_Y} \quad (10\text{-}2)$$

(see Held & Foley, 1994, for the derivation of this formula and its relation to the accuracy quotient Q derived by Gross, 1982). Gross showed that, when simultaneous

violations of the linearity and heteroscedasticity assumptions are such that ratio of the numerator value to the denominator are favorable, there can be an offset with an accurate estimate of the population validity. Because S_x is known, the ratio of b to S_y in Equation 10-2 becomes the determining index of correction accuracy. In fact, it is possible for *b (sample)* and S_y (estimated) to deviate slightly or wildly from the population values to produce an accurate validity estimate.

Assuming any specific sample size and population validity, would the standard deviation of a bootstrapped distribution of corrected validity coefficients be larger when there are large offsets in b and S_Y than when there are small offsets? The bootstrap distribution *SD* might be larger when there were large offsets because there would be the potential for more extreme pairs of non-offsetting b and S_Y values to be picked up through the "random selection with case replacement" bootstrap sample forming procedure. One could picture an upward curving pear as the example of an offsetting case with the slope and conditional *Y SD*s lowering as predictor scores lower.

Concluding Remarks

Violations in the assumptions for performing the correction for range restriction might or might not lead to an inaccurate estimate of the population validity. The idea that assumption violations can offset each other should be well understood. The personnel selection practitioner should consider understanding the sample as well as possible. A clinical approach could be used where a sample's adequacy is assessed by evaluating the consistency of the conditional prediction errors and regression weights across segments of the explicit selector score range, and also by studying the residuals from the sample regression analysis (y regressed on x). Of course, a clinical approach would require an adequate sample size and a selection ratio that is not too stringent. Even in the best of circumstances, we can never accurately extrapolate a sample b and s_y situation to the unrestricted population. The next chapter describes how small sample size and stringent selection ratios can result in wildly different range corrected validity estimates across samples.

Chapter 10. References

Brewer, J. K., & Hills, J. R. (1969). Univariate selection: The effects of size of correlation, degree of skew, and degree of restriction. *Psychometrika, 34,* 347-361.

Ghiselli, E. E. (1966). *The validity of occupational aptitude tests*. NY: Wiley.

Greener, J. M., & Osburn, H. G. (1980). Accuracy of corrections for restriction in range due to explicit selection in heteroscedastic and nonlinear distributions. *Educational and Psychological Measurement, 40,* 337-346.

Gross, A. L. (1982). Relaxing the assumptions underlying corrections for restriction in range. *Educational and Psychological Measurement, 42,* 795-801.

Gross, A. L., & Fleischman, L. (1983). Restriction of range corrections when both distribution and selection assumptions are violated. *Applied Psychological Measurement, 7,* 227-237.

Gross, A. L., & Fleischman, L. (1987). The correction for restriction of range and nonlinear regressions: An analytic study. *Applied Psychological Measurement, 11,* 211-217.

Guion, R. M. (1965). *Personnel testing*. New York: McGraw-Hill.

Held, J. D., & Foley, P. P. (1994). Explanations for accuracy of the general multivariate formulas for correcting for range restriction. *Applied Psychological Measurement, 18,* 355-367.

Holmes, D. J. (1990). The robustness of the usual correction for restriction in range due to explicit selection. *Psychometrika, 55,* 19-32.

Johnson, J. W., & Sager, C. E. (1991, April). *The robustness of range restriction correction due to incidental selection.* Poster presented at the Sixth Annual Conference of the Society for Industrial and Organizational Psychology, St. Louis, MO.

Lawley, D. (1943). A note on Karl Pearson's selection formula. *Royal Society of Edinburgh, Proceedings, Section A, 62,* 28-30.

Lee, R., & Foley, P. P. (1986). Is the validity of a test constant throughout the test score range? *Journal of Applied Psychology, 71,* 641-644.

Linn, R. L. (1968). Range restriction problems in the use of self-selected groups for test validation. *Psychological Bulletin, 69,* 69-73.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Pearson, K. (1903). Mathematical contributions to the theory of evolution - XI. On the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions of the Royal Society, London, Series A,* 200, 1-66.

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology, 85,* 112-118.

Chapter 11.
The Potential for a Negative Range Corrected Validity
Janet D. Held and Thomas R. Carretta

Introduction

From a Monte Carlo study reported in a later chapter, we will see that the distributions of validity coefficients corrected for range restriction could, in principle, include corrected coefficients with a negative sign when the population sign is positive. Ree, Carretta, Earles, and Albert (1994) provided a discussion with empirical examples regarding the potential for a corrected validity coefficient to change sign from its uncorrected condition. The issue arose when Air Force psychologists noted the sign change phenomena was sometimes observed for groups, particularly in pilot selection (Thorndike, 1949). The discussion initiated by Ree et al. (1994) is extended in this chapter with an examination of the basis for a sign change using conceptually simplified range restriction correction formulas (univariate and multivariate). The study data reported in this chapter were scores on the Armed Services Vocational Aptitude Battery (ASVAB) obtained from a Navy applicant population at a time when the battery contained the Numerical Operations and Coding Speed tests. A suitable criterion was designated as one of the 10 ASVAB tests so that population (unrestricted) validity coefficients were known. Some of the restriction in range formula derivations presented in previous chapters are repeated here for clarity.

How Negative Range Corrected Validities Can Occur

The general restriction in range problem (Pearson, 1903; Lawley, 1943) in military personnel selection research is to find the predictor/criterion correlation (validity) for the applicant population of interest, for which only predictor (selection instrument) information is available. On the basis of complete predictor/criterion information obtainable for a restricted subset of the applicant population (students selected at some predetermined minimum aptitude level), correction formulas can be used to estimate the unrestricted applicant population validity. As noted several times, the unrestricted applicant population is, theoretically, the one from which future recruits be selected for training school. The accuracy of this estimated (corrected) validity is contingent on the degree to which certain data assumptions have been met. These assumptions for the bivariate case of one predictor and one criterion are (a) linearity of regression of the criterion, y, on the predictor, x; (b) homoscedasticity of y error variance for all values of x; and (c) selection having occurred solely on x. The bivariate formula (correction for explicit selection) commonly encountered in the literature and cited in previous chapters (Case 1 from Guilford, 1965, p. 141; Case A from Thorndike, 1982, p. 210) is

$$R_{XY} = \frac{r_{xy}(S_X / s_x)}{\sqrt{1 - r_{xy}^2 + r_{xy}^2 (S_X / s_x)^2}} \quad . \qquad (11\text{-}1)$$

Equation 11-1 can be conceptually simplified in the more familiar form,

$$R_{XY} = \frac{S_{XY}}{S_X S_Y} \quad , \quad (11\text{-}2)$$

where R_{XY} is the corrected validity coefficient (not a multiple correlation), S_{XY} represents the unknown unrestricted population covariance, S_Y represents the unknown unrestricted population criterion standard deviation, and S_X is the known unrestricted population predictor standard deviation. S_{XY} is derived from the linearity identity where unrestricted and restricted slopes are assumed equal. S_Y is derived from the homoscedasticity identities where unrestricted and restricted standard errors of estimate across the entire range of the predictor are assumed equal. These slope and error identities, as we have seen in previous chapters are, respectively,

$$B = S_{SY} / S_X{}^2 = b = s_{xy} / s_x{}^2 \quad (11\text{-}3)$$

$$S_E{}^2 = S_Y{}^2 (1 - R^2{}_{XY}) = s^2{}_e = s^2{}_y (1 - r^2{}_{xy}). \quad (11\text{-}4)$$

Without solving the problem here (see Gulliksen, 1950 and Held and Foley, 1994 for formula applications using the ASVAB, as well as Chapter 5 of this document), the numerator in Equation 11-2 (the corrected covariance) presents the only possible opportunity for a negative sign. Further, from the linearity assumption, the corrected covariance is derived applying the restricted sample weight (unstandardized regression coefficient, or slope). Formally,

$$B = S_{XY} / S_X{}^2 = b = s_{xy} / s_x{}^2 \quad (11\text{-}5)$$

$$S_{XY} = bS_X{}^2, \quad (11\text{-}6)$$

where the sign of *b* determines the sign of both the restricted and corrected covariance, and therefore, the restricted and corrected validity coefficient.

The bivariate correction just reviewed is the singular case of the general multivariate correction (Gulliksen, 1950, Chp 13; Lawley, 1943, and Chapter 5 of this document). As described in Chapter 5, the multivariate correction formulas are merely matrix algebra extensions of the univariate case. The multivariate correction treating incidental selector variables as explicit is typically applied by military psychologists in an attempt to isolate all selection factors (Novick & Thayer, 1969) and because the procedure has been shown to be, generally, more accurate than the univariate correction (Booth-Kewley, 1985). The multivariate correction has also been shown to be more accurate with very large samples under violations of the correction assumptions and at stringent selection ratios, where inaccurate corrections are typically found.

A common correction assumption violation is heteroscedasticity where a flattening of the regression line is observed due to *y* variances typically being smaller for extreme values of *x* than in the middle range of *x* (Lord & Novick, 1968, p. 148). This study of small samples, however, reveals the inadequacies of the multivariate correction with small samples, where sampling errors in regression weights are high ("bouncing betas") exaggerated at very stringent selection ratios where there may be a flattening of the sparse data points.

Taking Equation 11-2 as the conceptually simplified correction formula for the bivariate case, we only need to generalize the multivariate covariance derivation parallel to S_{XY}. That parallel is in matrix algebra notation,

$$\mathbf{C}_{XY} = \mathbf{w}'_{yx}\mathbf{C}_{XX}, \quad (11\text{-}7)$$

where, for multiple predictors but only one criterion, $\mathbf{C}_{XY}$ and $\mathbf{w}'_{yx}$ are covariance and full least squares regression weight vectors, and $\mathbf{C}_{XX}$ is a square matrix of predictor (selector) variances/covariances. As in the bivariate case, $\mathbf{C}_{XY}$ is derived from linearity identities; however, the identities now apply to multiple selection variables selectors (e.g., ASVAB tests with known population parameters, are treated mathematically as explicit selectors even though operationally, the explicit selector may be a composite formed from only a few). Each covariance term, C_{XiY} in $\mathbf{C}_{XY,}$ is derived through matrix algebra.[9] This involves summing the multiplicative terms in two vectors: the particular selector test's variance with that selector's regression weight, and the subsequent covariances between that selector and every other selector multiplied by that other selector's weight. Given all selector variables are positively correlated, negative covariances in $\mathbf{C}_{XY}$ will be obtained through the matrix multiplication, if and only if, at least one weight is negative. And, there must be a sufficient magnitude or number of negative weights to produce the negative corrected covariance term (and thus, the negative corrected validity).

For predictors and criteria that are positively correlated in an unrestricted population, this unusual and theoretically impossible positive-to-negative sign change is a result of inadequate data. Next, we describe a study involving small samples and stringent selection to illustrate the unusual but theoretically possible case of a negative-to-positive sign change. All predictors and the criterion are positively correlated in this study's unrestricted population.

[9] Horst (1963) provides a clear presentation of matrix algebra for social scientists.

A Small Sample Simulation Study under Stringent Selection

The ASVAB selector composite, VE + AR (Verbal + Arithmetic Reasoning) was used as the explicit selection variable in a study examining the positive-to-negative range corrected validity coefficient phenomenon. VE + AR composite scores were used to select five random samples of 50 cases each from a Navy applicant population at the .10 selection (acceptance) ratio. The criterion was the ASVAB Mechanical Comprehension (MC) test, which was consistent with a prior study of Navy mechanical school classification composites (Held & Foley, 1994) where MC was used as a surrogate criterion but where the sample sizes were exceeding large.

Validities were corrected for range restriction in each of the five samples using both the univariate correction and two-predictor variable modified multivariate correction where (a) VE + AR was treated as the sole explicit selector and (b) VE and AR were entered separately into a multivariate correction as separate explicit selectors. Table 11-1 shows range corrected validities resulting from both correction procedures and the regression weights used in each correction. As both methods produced the same range corrected validities, they are listed only once.

Table 11-1
Uncorrected and Corrected Validities and Unstandardized Regression Weights used in Univariate and Modified Multivariate Corrections

		Univariate Case		Multivariate Case		
			Weights		Weights	
Groups	r_u	R_c	VE + AR	R_c	VE	AR
Unrestricted[a]	.687	.687	.451	.687	.425	.472
Restricted[b]	.184	.838	.601	.838	.549	.649
Sample 1 (*n* = 50)	.145	.759	.432	.753	.631	.274
Sample 2 (*n* = 50)	.092	.559	.301	.528	.483	.118
Sample 3 (*n* = 50)	.251	.917	1.004	.913	1.079	.912
Sample 4 (*n* = 50)	.121	.682	.332	.682	.395	.283
Sample 5 (*n* = 50)	.315	.938	1.031	.933	1.386	.765

Note: R_u is the uncorrected validity and R_c is the corrected validity, which was the same for each method. The composite VE + AR (Verbal + Arithmetic Reasoning) is the explicit selection variable.

[a]147,288 Navy applicants.

[b]13,684 Navy applicants selected by VE + AR at selection ratio = .10.

Table 11-1 only lists one set of range corrected validities for the five samples because the values were the same for both correction methods. None of the range corrected validities were negative nor were the test weights. The variation in the magnitude of the range corrected validities were tied to the magnitude of the uncorrected validities, which were subject to the small sample size (n = 50), the stringent .10 selection ratio, and the nature of the data at extreme segments of score distributions. Table 11-2 provides results for the eight-variable multivariate correction (the two explicit selection variables VE and AR augmented by the six remaining ASVAB incidental selection variables treated as explicit). Sample 4 of Table 11-2 shows a negative sign change in the corrected validity, which can be attributed to the large (erratic) negative weight for AR and its influence in the matrix algebra derivation of the AR covariance term (with MC).

Table 11-2
Uncorrected and Corrected Validities and Unstandardized Regression Weights used in Eight Variable Multivariate Range Corrections

			Weights							
Groups	r_u	R_c	**VE**	**AR**	**MK**	**AS**	**GS**	**EI**	**NO**	**CS**
Unrestricted[a]	.687	.687	.065	.224	.173	.319	.137	.173	-.050	.029
Restricted[b]	.184	.719	-.070	.226	.264	.247	.168	.207	-.040	-.008
Sample 1 (n = 50)	.145	.822	.392	.411	.354	.164	-.180	-.133	.280	-.194
Sample 2 (n = 50)	.092	.345	-.363	-.177	.164	.247	.578	.012	.095	-.058
Sample 3 (n = 50)	.251	.870	.383	.622	.066	.266	.363	.121	.112	-.172
Sample 4 (n = 50)	.121	-.316	-.291	**-1.044**	.527	.318	.537	.004	-.057	-.065
Sample 5 (n = 50)	.315	.878	.793	.479	.503	.608	-.237	-.134	-.063	-.199

Note: r_u and R_c are the uncorrected and corrected validities, respectively. VE + AR (Verbal + Arithmetic Reasoning) is the explicit selection variable. The other ASVAB tests applied as explicit selection variable in the correction are Mathematics knowledge (MK), Auto and Shop Information (AS), General Science (GS), Electronics Information (EI), Numerical Operations (NO), and Coding Speed (CS). Mechanical Comprehension (MC) is the criterion. Raw score weights were derived from a stepwise multiple regression procedure.
[a]147,288 Navy applicants.
[b]13,684 Navy applicants selected by VE + AR at the .10 selection ratio.

To explain the negative-to-positive sign change, we examined the three-variable correction case of one explicit selector variable (VE + AR [VEAR]) and two incidental selector variables (the criterion, MC, and a candidate replacement composite, VE + NO + CS [VENOCS]). The three-variable formula commonly encountered in the literature is

$$R_{ZY} = \frac{r_{zy} + r_{xz}\, r_{xy}\,[(S_x^2 / s_x^2) - 1]}{\sqrt{1 + r_{xz}^2\,[(S_x^2 / s_x^2) - 1]}\,\sqrt{1 + r_{xy}^2\,[(S_x^2 / s_x^2) - 1]}} \quad (11\text{-}8)$$

(Case 3 from Guilford, 1965, p. 343; Case C from Thorndike, 1982, p.213), where z is designated as the incidental selector composite, and x and y are designated as the explicit selector and incidental criterion variables, respectively. As with the criterion, population values for z are unavailable (at least treated mathematically so). As in the bivariate case, Equation 11-8 can be conceptually simplified to Equation 11-2 and further to the multivariate case using matrix notation. (see Horst, 1963, for matrix algebra applications for social scientists.) However, $\mathbf{C}_{XY}$, the corrected VE + NO + CS and MC covariance (individual composite test covariances summed for the composite covariance term), is taken from the $\mathbf{C}_{YY}$ matrix of derived incidental variance/covariance terms, as is the criterion standard deviation (square root of the diagonal variance term).

The potential for a negative-to-positive sign change for the incidental selection composite validity can be evaluated from the equation

$$C_{YY} = c_{yy} + w'_{yx}(C_{XY} - c_{xy}) \quad (11\text{-}9)$$

derived from the homoscedasticity identities.

Conceptually it is simpler to illustrate the inappropriate positive to negative range corrected validities graphically as in Figure 11-1 where the two predictors/selection instruments are not highly correlated in the unrestricted population.

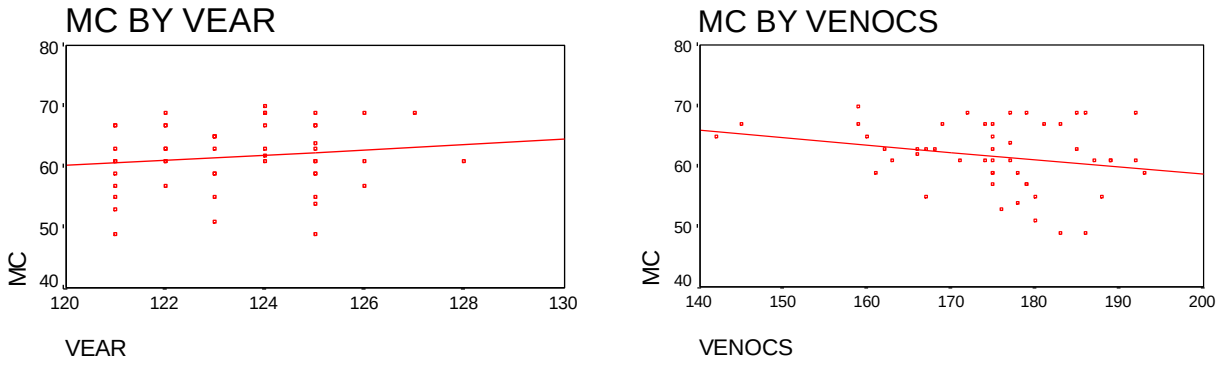


Figure 11-1. Bivariate predictor/criterion plot for the explicit and incidental selector variables.

We take from Figure 11-1 that complete truncation of the explicit selector at the stringent acceptance cutscore assures that at least a few high criterion outliers will exist for at least a few low-performing incidental selector scorers. Conversely, at least a few low outliers will exist for at least a few high-performing incidental selector scorers. If the incidental selector correlated highly with the explicit selector, the two graphs would be more similar. In fact, for this study, no negative restricted validity estimates were found for other composites that were more highly correlated with VE + AR.

The data used to generate the two graphs in Figure 11-1 are from Sample 1 of this simulation study. Note the stable regression weights for that sample (Composite scores are sums of standardized test scores: *M* = 50, *SD* = 10 in the ASVAB normative population). The restricted validity estimates of the explicit selector, VE + AR and the incidental selector, VE + NO + CS, are .145 and -.242, respectively. The restricted intercorrelation of the two selectors is .188. The unrestricted validity estimates for VE + AR and VE + NO + CS are .678 and .420, respectively. The unrestricted intercorrelation of the two selectors is .700. The explanation for the second graph and the obvious negative predictor/criterion relation stems from the rather low unrestricted validity of the incidental selector compared to the moderate validity of the explicit selector, and the moderate intercorrelation of the two selectors.

An Example Problem

The following is a simplified presentation of the multivariate correction for incidental range restriction for the three-variable case graphed above using the data from Sample 1 (see Table 11-2).

$\mathbf{C}_{XX}$ = [210.05] (unrestricted VE + AR *SD*); $\mathbf{c}_{xx}$ = [3.31] (restricted VE + AR *SD*)

$\mathbf{c}_{xy}$ = [1.43 3.78] (restricted VE + AR covariances with MC and VE + NO + CS)

$\mathbf{c}_{yy}$ = [29.53 -14.52] (restricted MC/VE + NO + CS incidental variable)

[-14.52 122.01] (variance/covariance matrix)

$\mathbf{w}_{xy} = \mathbf{c}^{-1}_{xx}\mathbf{c}_{xy}$ = [1/(3.31)][1.43 3.78] = [0.43 1.13]

$\mathbf{C}_{XY} = \mathbf{C}_{XX}\,\mathbf{w}_{xy} = \mathbf{w}/_{yx}\,\mathbf{C}_{XX}$ = [210.05][0.43 1.13] = [90.32 237.36]

$\mathbf{C}_{YY} = \mathbf{c}_{yy} + \mathbf{w}/_{yx}[\mathbf{C}_{XY} - \mathbf{c}_{xy}]$ = (corrected MC and VE + NO + CS variance/covariance matrix) = [67.75 85.92] [85.92 385.96], and thus the estimated population validity coefficient R_{XY} = 85.92 / (19.65 * 8.23).

The two denominator values, 19.65 and 8.23, were derived from C_{YY} as, respectively, the square roots of the MC and VE + NO + CS variance terms (in the diagonal). The corrected validity of VE + NO + CS is .530, which deviates from the actual value of .420 but is of the correct sign.

The Potential for Small Sample Stable Results: A Navy Study

It is not necessarily the fact that actual ASVAB multivariate corrected validity coefficients involving small samples at very stringent selection ratios will produce erratic regression weights as demonstrated in simulation studies involving multivariate normal distributions. In working with Navy schoolhouse grading systems, we note that much effort goes into the development of a high integrity criterion measure. Multiple progress tests are administered during the course of training that, for the most part, affect the final school grade. The progress tests and final test are developed to address the recruit abilities resulting from the ASVAB standard. In contrast, the underpinnings of the range correction formulas assume bivariate normality in the explicit selection two-variable case (x,y) and multivariate normality in the more than two variable case (e.g., x_1, x_2, y *or* x, y_1, y_2). Figure 11-2 illustrates the schoolhouse predictor/criterion situation where the full range ASVAB distribution is apparent, but not final school grade.

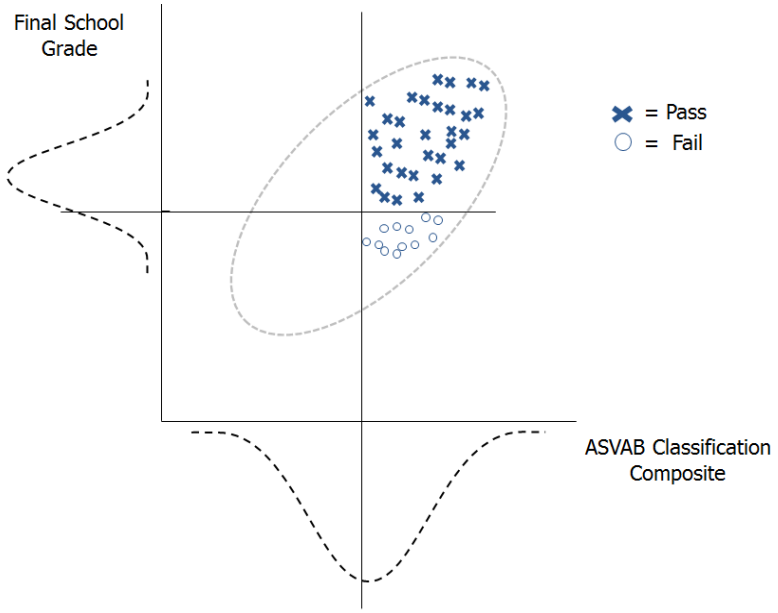


Figure 11-2. Bivariate predictor/criterion plot not fully mapped.

Figure 11-2 does not reflect a bivariate normal distribution in a full range population prior to selection even though the elliptical shape suggests it does. If the elliptical shape applied, the lowest scores of the final school grade distribution would stretch to the y-axis origin (as does the ASVAB score distribution on the x-axis). At this point in time, the Navy does not include school failure data (academically related or otherwise) in the validity analysis because (a) we may not have the exact reason(s) for failure, (b) the criterion data for graduates are considered high quality, and (c) the correction for range restriction equations would not know that the left tail of the y distribution is missing.

To illustrate the high integrity criterion data used in Navy ASVAB validation/ standards studies, Table 11-3 shows the ASVAB regression weights with final school grade as the criterion for two Navy Air Traffic Controller (AC Rating) A-School samples (initial technical training), each of very different sample size (n = 269 and n = 71) (Held, 2006).

Table 11-3
Unstandardized Regression Weights used in a Nine Variable Multivariate Range Correction for Two Navy Air Traffic Controller Training Samples

	Weights									
Sample Size	**GS**	**AR**	**MK**	**EI**	**AS**	**MC**	**VE**	**AO**	**CS**	**Sum**
***n* = 269**	.031	.113	.099	.031	.038	.073	.008	-.012	.132	.513
***n* = 71**	.086	.049	.248	-.095	.131	.006	.055	-.004	.166	.643

Unlike Table 11-2, which showed erratic regression weights for some of the small samples, Table 11-3 shows relatively stable weights, although different for each sample. Most importantly, the sums of the weights for the two samples are not highly dissimilar even though the sample sizes are. The sums of the regression weights across small samples, and not the signs of each variable in a sample, have been shown to be somewhat of an indicator of correction stability (Held and Foley, 1994). Table 11-4 lists the multivariate range corrected validities for a number of candidate ASVAB composites derived for each of the two AC samples whose regressions weights are listed in Table 11-3. (Coding Speed [CS] test is a former ASVAB test now a Navy special classification test.)

Table 11-4
Similarity of Multivariate Range Corrected Validities and Validity Differences for Two Navy Air Traffic Controller Training Samples

Predictor	**AC Rating Sample #1 (*N* = 269)**	**AC Rating Sample #2 (*N* = 71)**
VE+AR+MK+CS	.74 (largest)	.78 (largest)
VE+AR+MK+MC	.72	.75
AR+2MK+GS	.72	.76
VE+AR+MK+AO	.72	.76
VE+MK+GS	.70	.75
VE+AR	.67 (smallest)	.71 (smallest)

Note. Samples were taken several years apart (Held, 2006).

As can be seen from Table 11-4, both Navy Air Traffic Controller (AC Rating) A-School samples, taken several years apart, yielded rather stable multivariate corrected validity results despite the disparity in their sample sizes (n = 269 vs. n = 71). The largest validity coefficient for each sample was for the VE+AR+MK+CS composite (.74 and .78 for the larger and smaller samples, respectively) and this stability in results was related to the stability in the sum of the regression weights (Table 11-3).

Concluding Remarks

This chapter described several conditions under which sign changes can occur when correcting validity coefficients for range restriction using the multivariate method. In general, the negative-to-positive sign change when all selector variables and the criterion are positively correlated in the unrestricted population is a function of the intercorrelations of the selectors and criterion in the restricted data set and cannot be viewed as an abnormal outcome. The positive-to-negative sign change may merely be due to a highly stringent selection ratio combined with a small and/or inadequate data set and should be viewed as an unrealistic outcome.

Even though the methodological issues revealed in this chapter's study and Chapter 10 on assumption violations, we may not have to be overly cautious about applying the multivariate correction formulas if the samples are not extremely small because of the Navy's high integrity process for developing the training performance criterion. Also, whereas the selection ratio in the simulation study was extremely stringent (top 10% qualification rate), they are much more moderate for most of the Navy's (and other Services') military occupations.

The next chapter departs from the range correction topic and addresses several topics that are commonly considered when estimating the relation between a selection or classification instrument and a performance criterion.

Chapter 11. References

Booth-Kewley, S. (1985). *An empirical comparison of the accuracy of univariate and multivariate corrections for range restriction* (NPRDC-TR-85-19). San Diego: Navy Personnel Research and Development Center.

Guilford, J. P. (1965). *Fundamental statistics in psychology and education*. NY: McGraw-Hill.

Gulliksen, H. (1950). *Theory of mental tests*. New York: John Wiley & Sons.

Held, J. (2006). *Armed Services Vocational Aptitude Battery (ASVAB) standards: Air Traffic Control rating* (NPRST Letter Report Ser 3900, PERS-1/00047 31 May 2006): Millington, TN: Navy Personnel Research, Studies, and Technology.

Held, J. D., & Foley, P. P. (1994). Explanations for accuracy of the general multivariate formulas in correcting for range restriction. *Applied Psychological Measurement, 18,* 355-367.

Horst, P. (1963). *Matrix algebra for social scientists*. NY: Holt, Rinehart and Winston, Inc.

Lawley, D. (1943). A note on Karl Pearson's selection formula. *Royal Society of Edinburgh, Proceedings, Section A, 62*, 28-30.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Novick, M. R., & Thayer, D. T. (1969). *An investigation of the accuracy of the Pearson selection formulas* (ONR-RM-69-22). Princeton NF: Educational Testing Service.

Pearson, K. (1903). Mathematical contributions to the theory of evolution - XI. On the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions of the Royal Society, London, Series A,* 200, 1-66.

Ree, M. J., Carretta, T. R., Earles, J. A., & Albert, W. (1994). Sign changes when correcting for range restriction: A note on Pearson's and Lawley's selection formulas. *Journal of Applied Psychology, 79*, 298-301.

Thorndike, R. L. (1949). *Personnel selection*. New York: Wiley.

Thorndike, R. L. (1982). *Applied psychometrics*. Boston: Houghton Mifflin Company.

Chapter 12.
Partial Correlation, Hierarchical and Logistic Regression, and Power
Thomas R. Carretta and Janet D. Held

Introduction

This chapter addresses several topics that are commonly considered when estimating the relation between a selection or classification instrument and a performance criterion: (a) partial correlation to remove a variable's influence, (b) hierarchical regression to estimate a variable's influence (e.g., variables fixed in the study such as demographics, or variables whose measures have been taken at different points in time), (c) logistic regression when continuous criterion measures are not available, and (d) power analysis. We note that non-linear relations are not addressed in this manual, but there are methods fully addressed in the literature.

Partial Correlation: The Effect of a Third Variable

Validation research is generally correlational in nature. The interpretation of correlations, although straightforward on the surface, can be fraught with hazards. Consider the correlation of an ability test with supervisor ratings of job performance. It would not be unusual to find low correlations, which could lead to inappropriately abandoning predictive measures. The *Principles for the Validation and Use of Personnel Selection Procedures* (SIOP, 2003) noted that the relation between ability (or any other measure) and occupational criteria is best understood with the effect of job experience removed. That is, those individuals who have prior experience with performing the job tasks, or tasks that are similar to those involved in the current job, will naturally perform at higher levels, at least in the beginning, all other things being equal. Removing, or controlling for, this "experience" variable can easily be done by "partialing out" experience from the relation between ability and the criteria. Partial correlation, in a more general sense, measures the degree of association between two variables with the effect of one or more variables removed. The partial correlation is computed with the following formula:

$$r_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_1} r_{x_1x_2}}{\sqrt{1-r_{yx_1}^2}\sqrt{1-r_{x_1x_2}^2}}, \qquad (12\text{-}1)$$

where $r_{yx_2 \cdot x_1}$ is the correlation between y and x_2, while partialing out the effects of x_1; r_{yx_2} is the correlation between y and x_2; r_{yx_1} is the correlation between y and x_1; and $r_{x_1x_2}$ is the correlation between $x_1$and x_2 (Crocker & Algina, 1986). Using our example, y is job performance ratings, x_2 is the ability test, and x_1 is job experience.

Carretta, Perry, and Ree (1996) provided an example when they correlated ability test scores with ratings of situational awareness (SA) for 171 F-15 pilots. The zero-order correlation (zero-order is the term used to indicate that no partialing out has been done) of ability and SA was .10. However, when F-15 flying experience (i.e., number of flying hours) was partialed out, the correlation was .17. In this instance, it would be incorrect to report the correlation between ability and SA as .10.

More broadly, the idea of partial correlation can be subsumed under the statistical concept of mediation. Mediation means that one variable acts through another to exert its influence on a third variable. For example, "A →B→ C" indicates that variable A acts through variable B to exert its influence on variable C. Note that there is no *direct* influence of A on C in this model specification. That is, we do not specify "A→C" although that relation can occur.

Hunter (1986) provided an informative model of mediation in the area of job performance. Hunter demonstrated for numerous jobs that job knowledge mediated the relation between ability and job performance. Similarly, Ree, Carretta, and Doub (1998/1999) showed for 83 U.S. Air Force enlisted jobs that prior job knowledge (JK_P) mediated the relation between ability (the general ability *g* factor) and the acquisition of subsequent job knowledge (JK_S) during training. In this case both ability and prior job knowledge were directly related to the acquisition of subsequent job knowledge during training, but ability also had an indirect influence on subsequent job knowledge through prior job knowledge. The path diagram is shown in Figure 12-1.

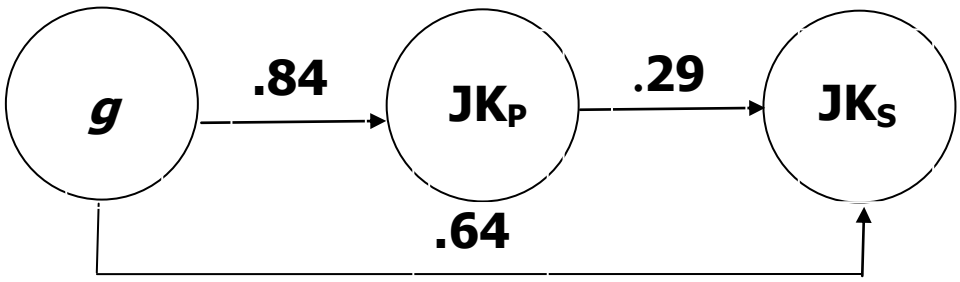


Figure 12-1. An example of mediation: Ability (*g*) and prior job knowledge (JK_P) have a direct effect on the acquisition of subsequent job knowledge (JK_S); ability also has an indirect effect on JK_S through JK_P (Ree et al., 1998/1999).

The effect of mediation also was demonstrated by Ree, Carretta, and Teachout (1995) for pilot trainees. In the Ree et al. (1995) study, general cognitive ability (*g*) had both direct and indirect influence on the acquisition of aviation job knowledge and on hands-on flying performance during pilot training. Several other studies have shown the mediating effect of job knowledge between ability and performance (Borman, White, & Dorsey, 1995; Borman, White, Pulakos, & Oppler, 1991; Lance & Bennett, 2000; Schmidt, Hunter, & Outerbridge, 1986).

In closing, to know the "true" relation of a predictor with job performance (assuming perfect reliability) it is necessary to partial out the effect of experience, such as job experience (i.e., years, training). The variable to be partialed out, however, depends on the purpose of the study. Most importantly, partialing out the effects of a variable should be considered when the objective is to confirm one's theory about the variables that contribute to performance. One possible military application of the partialing, or control, procedure would be for reclassifying enlisted members when they are required to do so (e.g., due to military downsizing or reprioritization in staffing occupational fields). One could control for the number of years in service when examining the relation between ASVAB scores and re-training outcomes. For example, the Air Force takes into account the experience of pilots manning aircraft when selecting candidates for training on Remotely Piloted Aircraft (RPA). In this case, the control variable would be number of flying hours.

Hierarchical Regression

One of the Navy's analysis tools applied in some ASVAB validation/standards studies is hierarchical regression. This tool, or method, is used primarily to evaluate the validity contribution of existing or candidate tests that serve as a second stage classification screen (multiple hurdle). Hierarchical regression analysis is also a useful method for controlling for the effects of variables, like demographic variables, that are hypothesized to relate to the dependent variable of interest but that are not in realm of control (i.e., a variable available in the dataset but not a variable manipulated via an experimental design or considered a suitable basis for rejecting candidates). Unlike stepwise regression where the variable accounting for the most variance in the criterion variable automatically enters into an equation first, and so forth, until a single model is developed, in hierarchical regression, the researcher determines the order of variable entry. Each entry step produces a regression equation and associated statistics, building up the models to account for the added variables at each step.

We refer to hierarchical regression as linear in this section with variable entry sequentially stepped so that the multiple R^2 and change in R^2 can be evaluated for every added variable. Note in this case that "R" in fact refers to a multiple correlation whereas in past chapters, aside from a specific application in Chapter 10, "R" stood for the unrestricted (population) validity coefficient. Note also that non-linear models may apply that may have complicated interpretations due to unaccounted multiple hurdle selection systems (that we do not address here).

In the Navy ASVAB validation/standards context, hierarchical analysis has been used to assess whether the Defense Language Aptitude Battery (DLAB) contributes validity above the ASVAB in predicting final school grade in language training. The method also has been used to assess whether the Nuclear Field (NF) community's Navy Advanced Placement Test (NAPT) contributed validity above the ASVAB in predicting final school grade in highly technical training courses. Both of these tests are expensive to develop, update, administer, and maintain (not to mention examinee testing time, as each test can take 2 hours to complete).

We refer the reader to the Nuclear Field (NF) study (Appendix B of the Introductory Manual) for an example of a hierarchical regression analysis. Briefly, the analysis was conducted for the three NF Ratings (Electronics Technician, Electrician's Mate, and Machinist's Mate) (Table 8 in Appendix B) to determine the significance of the NAPT's incremental validity to the ASVAB. NAPT was entered as a second step in the hierarchical analysis after the ASVAB and the incremental validity of the NAPT was statistically and meaningfully significant for all three Ratings. Also significant was a subsequent entry step, waiver type (e.g., education), and for two Ratings, the following regression step in the hierarchy, the interaction term of two different types of waivers (education and civil).

We refer the reader to Lautenschlager and Mendoza (1986) for the type of hierarchical analyses described in this section and to Raudenbush and Bryk (2002**)** for a complete treatment of hierarchical/multilevel analysis. We close this section by noting that partial correlation and hierarchical regression (linear in both cases) essentially address the objective of controlling for variables or isolating their effects from other variables and that the "experience variable" discussed in the partial correlation section could just as well have been entered into a hierarchical regression analysis. The difference is the focus of the analysis: eliminating an influence or establishing its importance. Power analysis (Cohen, 1988) can be applied to both methods and also to logistic regression, which is discussed in the next section.

Logistic Regression when the Criterion is Binary

Raju, Steinhaus, Edwards, and DeLessio (1991) suggested that a two-parameter logistic regression (LR) model could be used for several personnel functions that involve selection instruments, most notably in setting cutscores conditional on ability. The outcome of interest in this application of LR is not the magnitude of the validity coefficient but the predicted probability of success (e.g., passing training) conditional on the selection instrument scores (e.g., the ASVAB). Raju et al. noted the usefulness of LR in that the validity coefficient that results from the correlation of two continuous measures in isolation does not have practical value in setting cutscores (or assessing utility for an organization – see Chapter 3) unless used with the Taylor-Russell tables (Taylor & Russell, 1939). We note, however, that the Taylor-Russell tables are an integral part of conducting ASVAB validation/standards studies (discussed earlier in Chapter 3). We are reminded that these tables allow us to estimate expected improvements in military training success rates conditional on validity magnitude, selection stringency, and current success rates under different scenarios. The interest here is in aggregate success rate of individuals in a training course, not how likely it is that any individual will pass.

The LR predicted probability of success (or failure) is not without merit or place in ASVAB validation studies, but only when the performance outcome is not measured on a continuous score scale. For example, the ASVAB validation/standards study conducted for the Navy SEALs (Sea, Air, and Land special warfare combat forces) provided in Appendix A of the Introductory Manual illustrates the use of LR when a continuous criterion variable is not available, just a pass/fail binary outcome. The

SEALs conduct mentally and physically challenging training and there are many reasons for not completing training (e.g., medical issues that arrive in training). The performance measure to date is not scored on a continuous scale. The most frequent reason for not passing SEAL training, at least in the initial BUD/S course, is self-elimination (drop on request – cannot meet the physical or mental challenge).

The SEAL ASVAB study was an attempt to establish which of a number of ASVAB composites was most predictive of a successful training outcome comparing the LR procedure with the Lawley (1943) multivariate correction for range restriction (discussed in Chapter 5 and elsewhere). The validity coefficient rankings (magnitude) among a number of candidate ASVAB composites were compared across the two methods recognizing that the (a) the binary criterion variable suppressed the validity coefficient in each case and (b) only the Lawley procedure would adjust for the downward bias of the validity coefficient due to ASVAB selection effects (SEALS had two alternative standards at the time) without the correction (i.e., the LR procedure).

In LR, there are several reported pseudo validity coefficients, one of which was used in the validity coefficient comparison, the Nagelkerke *R* (square root of the reported Nagelkerke R^2). The Nagelkerke R^2 is recognized as inappropriate for comparison to the OLS R^2 on theoretical grounds but perhaps useful in comparing models (Hosmer & Lemeshow, 2000). Our goal in the SEAL study was just that – that is, for each method (multivariate range correction vs. LR), to determine whether the ASVAB composites were ranked the same, which they were, but with lower validity magnitudes for LR.

We note that another shortfall in the comparison study was that the corrected validity coefficient obtained from the Lawley procedure was not really a multiple *R,* but rather an estimated Rho for a particular integer-weighted ASVAB composite. The Ordinary Least Squares (OLS) equation weights (that result in a multiple *R* in the sample) are merely applied in the multivariate range restriction correction (Chapter 5).

The magnitude of the "population" validity coefficient is important in ASVAB validation/standards studies, and when a binary (pass/fail) score implies an underlying continuous score distribution, a correction for this artificial dichotomization (not applied in the SEAL study) should be applied, as described in the next section.

Correction for Dichotomization: An Alternative Approach to Logistic Regression When the Criterion Is Binary

LR is most appropriate when the criterion variable is a genuine dichotomy, such as "crashed the airplane" versus "landed safely," or most commonly in the study of disease interventions where the patient either died or lived. The question of whether a student passed or failed, however, implies an underlying performance distribution where a decision point or cut-point on the final school grade establishes how the student is categorized. Did the computer technician pass or fail the information network certification test? A "yes/no" answer may be the only type of available performance information. When the grades that determined pass/fail disposition are available, they should be the criterion measure of choice and we then deal with range restriction to estimate the validity coefficient (past chapters).

Logically, we can think of an individual who barely passed the course as not being as knowledgeable as one who passed with flying colors. In the Navy SEAL ASVAB study (Appendix A in the Introductory Manual), the training disposition available was either pass the grueling mentally and physically challenging regimen, or drop on your own accord. We could think that even the dropped students had some differentiation in "What it takes to be a SEAL", but there is no measurement instrument used to differentiate students who drop.

Dichotomization of a continuous criterion variable (yielding an *artificially* dichotomized variable, as opposed to a truly dichotomous variable) not only loses information, but causes the correlation between it and a continuous variable to appear lower than it theoretically should be. If we compute a simple correlation between a continuous predictor variable (say a cognitive test score) and a dichotomized dependent variable (say pass/fail in a training course), we are computing a point-biserial correlation. (For SPSS/SAS users, the point-biserial is merely the Pearson correlation between the continuous variable and the dichotomized variable – artificially dichotomized or a true dichotomy.)

We learn in basic statistics that when the proportions are 50-50 for each category of the dichotomized variable (e.g., 50% pass and 50% fail), the variance is maximized relative to other splits. Because variance is maximized, so is the correlation coefficient. The more extreme the splits, the more downward is the biasing effect on the correlation. For example, if the correlation between two variables is .50 before dichotomization and the proportions are 50-50 in the dichotomized criterion, the correlation after dichotomization will still be .50. However, if the proportions are 60-40, 70-30, 80-20, and 90-10, the correlations from dichotomization will be .39, .38, .35, and .29, respectively. If the correlation before dichotomization is .25, the after-dichotomization correlations for the proportions 60-40, 70-30, 80-20, and 90-10 would be .20, .19, .17, and .15, respectively. We see that the lower the correlation to begin with, the lower the downward biasing effect when one of the two continuous variables is dichotomized.

Cohen (1983) reminds us that early test construction practices (before the computational efficiency that computers provide) involved dichotomizing the total test score into high/low (analogous to pass/fail for training) to simplify the computation of test item/total test score correlations. Cohen did not object to this practice when the only purpose was merely to decide which items to keep and which to toss. From Cohen's article, the rank ordering of items based on the point-biserial correlation does not appear to be affected by this total test score dichotomization. However, to correct for the correlation underestimate (assuming an underlying continuous distribution), Cohen (1988) and Cohen and Cohen (1983) provide the statistical correction obtaining the biserial from the point biserial, which involves the ordinate (height) of the normal curve at the point of dichotomization (assuming a normal distribution of underlying scores/abilities/etc.).[10]

[10] The ordinate value (the height of the normal distribution at the cutpoint) is available in some statistical appendices, including Cohen and Cohen (1983) Appendix Table C.

The Air Force has long recognized the dichotomization problem and potential ways to deal with the complication of restriction in range. The Air Force applies the correction for dichotomization (e.g., when pass/fail in training is all that is available) after correcting validity coefficients for range restriction in order to provide a better estimate of the predictive validity of personnel selection methods. The Navy has not applied the procedure only because, at this point, only the Navy SEALs have been found to only have a dichotomous training outcome variable. We illustrate the correction for dichotomization (that is, obtaining the biserial correlation) subsequent to a multivariate range correction for a .34 ASVAB validity reported in Table 6 of Appendix A of the Introductory Manual (SEAL study):

$$r_b = r_{pb}\frac{\sqrt{pq}}{h},$$

where r_b is the Biserial correlation and r_{pb} is the Point Biserial, in this case, corrected for multivariate range restriction (.34), pq = .30 X .70 (30% pass rate for SEALs) and h = .352 taken as the height of the normal curve (symmetrical for .30 and .70) from Appendix C in Cohen and Cohen (1988). The r_b value turns out to be .443. Of course it will never be known whether these two corrections (for range restriction and dichotomization) yield an accurate estimate of the ASVAB's validity in the ASVAB normative youth population (PAY97) in predicting SEAL training outcomes.

Finally, Cohen (1983) cautions us against artificially dichotomizing the criterion variable because not only is the correlation coefficient diminished, but so is statistical power, a topic discussed in the next section.

Statistical Power

Background

Statistical power is the probability of detecting a statistically significant difference in a sample when in fact it exists in the population. More formally, power is the probability of rejecting the null hypothesis (H_o) when it is false and therefore accepting the alternative study hypothesis (H_1) when it is true (Cohen, 1988). Almost all statistics courses include the topic of statistical power (Cohen; Cohen & Cohen, 1983), but relatively few published studies report power to accompany the various statistical tests to which power can be applied(e.g., r, t, Z, or F tests). Two surveys of a prestigious applied psychology journal showed that the average statistical power for studies accepted for publication was only .46 and declined to .37 two decades later (Sedlmeier & Gigerenzer, 1989). In other words, researchers could only expect to detect an existing effect 46% of the time (37% for the more recent survey). Conversely, researchers could expect to *fail* to detect the existing effect 54% of the time (63% for the more recent survey)!

Low statistical power is not unique to psychological research and has been reported in reviews of management research (Cashen & Geiger, 2004; Mazen, Graf, Kellogg, & Hemmosi, 1997), software engineering (Dyba, Kampenes, & Sjoberg, 2006), and other fields such as medicine (Halpern, Karlawish, & Berlin, 2002). Low statistical power in the field of personnel selection research means that we are inclined to make incorrect conclusions about the psychological phenomena we study. In the case of the military's ASVAB validation/standards studies, an incorrect decision in recommending an ASVAB candidate composite for classifying recruits into a particular occupation would not have a dire consequence-because there already is an operational ASVAB composite with a cutscore in place (with the exception of newly formed occupations). Further, the candidate ASVAB composites evaluated for an occupation are generally highly correlated due to the rational composite development process that maps the underlying ASVAB constructs to the curriculum (although there is also an empirical approach that might produce less highly correlated composites).

In contrast, there can be a lot at stake in falsely rejecting a potentially effective medicine when the effect is real in the population. For example, making a false decision to *not* submit a drug to the FDA for approval because of low power may have substantial negative impact on those in the population who would have benefited from the drug. We stress that just because an effect is not detected in a sample taken from a population does mean the effect does not exist. In ASVAB standards validation work, sample size has a lot to do with detecting a real effect and it, out of all the other influential factors (including development of best practice criterion measures), may be the only variable within the practitioner's control.

Power analysis involves many moving parts. As Cohen (1988) notes, statistical power is a *joint* function of the Type I error rate, effect size, sample size, and degree to which the sample values reflect their true values in the population. We refer the reader to Cohen for the many power applications and merely review some fundamentals that apply in general. Researchers set a statistical significance level, alpha (α), for rejecting the null hypothesis, H_o, and thereby accepting the study hypothesis, H_1. Generally, a .05 α level in personnel psychology research is sufficiently stringent for us to reject the null hypothesis (H_o) and accept the alternative hypothesis (H_1). If the effect *does* exist in the population, increasing α will increase the probability that we detect that true effect (e.g., α = .10 rather than a more stringent .05) and thus increase statistical power. However, in sampling theory, we are bound to observe statistical significance some proportion of time over the long haul even when the effect is not there. The probability of doing so is called a Type I error and *is* the α level.

Establishing a more stringent statistical test (e.g., α = .01 rather than .05) reduces the Type I error but increases the probability of a Type II error (β), given the effect exists in the population. A Type II error is the probability of failing to reject H_o when in fact, the effect exists in the population and H_1 should be accepted. The cost of committing each type of error is weighed by the organization conducting the research. Of particular concern in the practical world is that extremely large samples will demonstrate an effect when the magnitude of that effect will have trivial consequences (Murphy, Myors, &

Wolach, 2009). Murphy et al. suggested considering a "minimum" effect size in determining sample size requirements in power analysis.

The calculation of power is simply 1 minus the probability of failing to appropriately reject H_o (Type II error β), or 1 – β. Increasing power is mainly about reducing β by limiting the overlap of the two distributions that are being compared for the effect. Diminishing the overlap can be achieved, basically, in three ways. First, increasing α reduces distribution overlap by moving the critical significance test value (in effect, a vertical cutscore) to the left on the *x*-axis. (We are assuming that the control group's distribution is to the left of the experimental group's distribution, with some overlap in the two). Second, moving the two distributions apart reduces overlap, which means somehow increasing the effect. Third, increasing sample size reduces overlap by narrowing the two distributions' spread (variance), thereby reducing overlap.

Power's Relevance for ASVAB Test Validation

We have so far stressed the importance of the following factors in ASVAB validation/standards studies: (a) applying the multivariate correction for range restriction in estimating ASVAB composite validity coefficients in the unrestricted ASVAB population, (b) assessing (to the extent possible) the sample's adherence to the underlying assumptions of performing the correction, (c) using (if not proactively helping to develop) meaningful and reliable performance criteria, and (d) having an adequate sample size to produce more accurate and stable range-corrected validity estimates. Of these factors, only adequate sample size pertains to the traditional calculations presented in power analysis.

Four questions logically arise about adequacy of sample size in conducting ASVAB validation/standards studies:

1. Is the sample at hand considered representative of the ASVAB population from which it theoretically was drawn; thereby leading us to believe that an effect found in the sample generalizes to that population?
2. Should we rely on historical studies in considering the actual effect size in the population (correlation coefficient/validity)?
3. If we refer to historical studies for a "pre-estimate" of validity magnitude and therefore sample size requirement, do we refer to the restricted or unrestricted ASVAB validity?
4. What magnitude of the validity difference should be considered sufficiently large to recommend one ASVAB composite over another?

Regarding the first question, the sample at hand is all we have, and we must therefore assume that it is representative of the ASVAB population. The question becomes *which* ASVAB population and whether the choice influences which parameter values (means, standard deviations, and correlations) are appropriate to apply in the range correction procedure. The Navy applies the PAY97 ASVAB population in the range correction procedure, so it assumes representativeness. We could decide that the most

recent military applicant population (or the Service-specific applicant populations) is closest in attributes to our sample. However, this produces at least two unintended consequences: (a) the inability to generalize validity coefficients over time due to changes in economic conditions that may lead to differing applicant compositions and (b) the inability to generalize validity coefficients across the Services' same occupations, which might be an issue under budget constraints that lead to more joint-service training and operations.

Regarding the second question, the American Psychological Association (APA) Task Force on Statistical Inference (Wilkinson, 1999) endorsed the approach of referring to high-quality historical studies for pre-estimating population effect sizes in power analysis. Regarding the third question, if we do refer to historical studies, do we refer to the restricted or estimated unrestricted ASVAB validity? If we refer to the estimated unrestricted validity, the ASVAB validation/standards study researcher must first complete the study—not at all helpful in the preplanning stages that involve estimating sample size. Regarding the fourth question, the comparison of ASVAB composite validity coefficients must involve the range-corrected state. Otherwise, the comparison will be biased (as we saw in the Chapter 5, Thorndike example - Table 5-1).

Determining the necessary sample size for sufficient power in an ASVAB validation/standards study is a complicated matter, and we look to others for advice in this area. We first look to regression analysis, because all of the ASVAB tests are used in sample-based regression weights that are then applied (via the linearity assumption along with the homoscedasticity assumption) in the derived multivariate range-corrected validity estimates. Green (1991) reviewed various "rules-of-thumb" presented in the literature regarding multiple regression and noted that the most conservative sample sizes came from Nunnally (1978). Nunnally applied a multiple regression shrinkage formula (p. 180) in his sample size analysis and recommended 300 to 400 cases for 9 or 10 predictor variables. Green (1991) took the position that if power is to be considered, the more complicated "rules of thumb" (that he reviewed) produced lower sample size requirements, but there was no specific endorsement at the time of a most appropriate integrated power/regression procedure.

Because the Navy applies all nine ASVAB tests in the multivariate correction for range restriction, we might consider a reasonable regression-based sample size estimate based on Nunnally's (1978) advice of 35 cases X 9 variables = 315 cases as a general rule. We would have to recognize, however, that we need a good understanding about factors that might influence regression weight stability (as described in previous chapters). We could also posit that being able to detect a real difference in the validity of ASVAB composites in the population is almost as important as detecting that the composites have a certain magnitude of validity (effect size). Cohen (1988) addressed both of these magnitude concerns, stating that "these (statistical power) tables are not valid under conditions of range restriction such as may occur in personnel selection" (p. 100). However, we could stretch our perspective a bit and refer to Cohen's Table 4.3.2 to test whether a .03 validity difference could be detected with sufficient power given different sample sizes, different validity magnitudes, and different validity differences between, say, two ASVAB classification composites.

Cohen's (1988) Table 4.3.2 (one-tailed α = .05 significance test) applies to the population effect size. However, the two correlations to compare are independent, coming from different populations (pp. 119-120). This is not the case in our ASVAB validation/standards studies; nevertheless, we proceed. The effect sizes of the difference are expressed as "*q*" and are listed across the top ledger row. The "q_c" values listed in the table column just to the right of the first column (with *N*s) apply if we are conducting significance testing for the correlations obtained from our samples. A further modification can be made for the case of paired correlations compared in one sample (section 4.5.4, p. 142).

For simplicity and illustrative purposes, we use Cohen's (1988) Table 4.3.2 to demonstrate a power planning scenario that would be useful to us if not for the restriction in range problem. We also note that the table applies only to correlations derived in independent samples (which apply to independent populations). Cohen specifies the effect size of the difference in correlation coefficients as "*q*", which is calculated as the difference in the Fisher *Z* transformation values calculated for each of the comparison correlations. Cohen assumes small, medium, and large effects sizes for correlations are associated with *q* values of .10, .30, and .50 (p. 129).

We first refer to Cohen's (1988) Table 4.2.1 after deciding that a small effect (*q* = .10) is sufficient to recommend one ASVAB composite over another. With simple calculations, we can determine the correlation differences needed to attain *q* = .10 for three baseline correlations that we specify as representative of the ASVAB validity for a range of Navy occupations. The three correlations are .30, .55, and .80, and the increments in validity required for *q* = .10 are .08, .05, and .03, respectively. We see right away that our specified .03 validity difference at the planning stages of the study deemed acceptable for an ASVAB composite replacement does not apply across the spectrum of validity magnitudes (i.e., if we were to assume the unrestricted validity coefficients for ASVAB composites were the target validity values to use to conduct the power analysis). More specifically for the reader, "*q*" is calculated from Cohen's (1988) Table 4.2.1 for the following comparisons: (a) .80 compared to .83 for a .03 difference (similar to our Nuclear Field study reported in Appendix B of the Introductory Manual), (b) .55 (Navy average) compared to .60 for a .05 difference, and (c) .30 (a bit larger than observed for the SEALs) compared to .38 for a validity difference of .08. The *q* values for these values are all slightly below .10 (.089, .075, and .090 for the baseline .80, .55, and .30 correlations, respectively).

Assuming the *q* values derived for our three correlation pairs are close enough to Cohen's (1988) *q* = .10 value, we enter Cohen's Table 4.3.2 at the .10 *q* column and look to see what sample size is required to achieve the power level we specified in the ASVAB validation/standards study planning stages (power of .80). A sample size of 1,000 is required to achieve the highest power level entered in the table entry (.72). We note that validity differences of as large as .08 (for the .38 - .30 comparison) are never observed in ASVAB validation/standards studies because of the substantial correlation between ASVAB composites that occurs through the rational composite development process of mapping ASVAB constructs to the curriculum constructs (complementing the empirical regression-based process).

We remind ourselves again that the power/sample size analysis we conducted using Cohen's (1988) Table 4.3.2 is inappropriate for ASVAB use in that the table applies to independent correlations, not related correlations derived in the same sample. We refer the reader to Cohen's modification for a pair of correlations (p. 142) and note that Equation 4.5.6 applies for testing the statistical significance (once the sample values are found) as opposed to research planning purposes.

Next, we look to others who have explicitly considered the range restriction issue. Schmidt, Hunter, and Urry (1976) were concerned with the legal issues surrounding use of a selection instrument with zero validity and so addressed the role of power in the personnel selection framework. Schmidt et al. acknowledged the complication involved in violating the assumptions underlying range correction and so conducted their research under the assumption that all had been met. The authors addressed explicit selection only, acknowledging that the incidental selection case is also important. In this regard, Sackett and Wade (1983) extended the Schmidt et al. power formulas and tables to include the incidental selection case, which they recognized as the *typical* personnel researcher's case of interest (i.e., evaluating new measures). We direct the reader to Raju, Edwards, and LoVerde (1985) for their comments on both articles.

We refer only to Schmidt et al. (1976) because the Navy treats all ASVAB tests as explicit selection variables in the multivariate correction for range restriction (applying all ASVAB regression weights derived in the sample). Schmidt et al. melded power, restriction in range, and criterion unreliability into their complex of equations and derived restricted validity estimates for 10 selection ratios and 8 criterion reliability estimates. Again, the objective for their personnel selection issue was only to establish if the observed validity was greater than zero in the unrestricted population. Schmidt et al. provided tables with sample size requirements for two unrestricted validity levels (.35 and .50), two power levels (.50 and .90), and various significance levels for one- and two-tailed statistical tests.

We refer to Schmidt et al.'s (1976) Table 4 that applies to the unrestricted validity of .50 (close to the Navy average of .55 across Ratings), a power level of .90 (higher than the Cohen's .80), a one-tailed $\alpha = .05$ test, and a criterion reliability of .80 (perhaps reasonable to assume across Navy schools). At a fairly unrestricted selection ratio of .70, a known unrestricted validity of .50 is reduced to .33 with a sample size requirement of 75. At a much more stringent SR of .30, the unrestricted validity of .50 is reduced further to .25 with a near doubling of the sample size requirement to 134. If we specify the .50 power level (flip of a coin) commonly found in cited historical literature and seemingly unacceptable in the personnel selection realm, then the sample size requirement is substantially reduced (*N*s = 26 and 44, respectively).

We close this section by saying that it is not clear in our minds what methods are appropriate for determining the sample size requirement for a specific ASVAB validation/standards study. The situation is complicated by practical issues such as the (a) urgency of the standards requirement that might preclude waiting for a sufficient sample size to form; (b) ability to detect obvious idiosyncrasies of the sample that might lead us to wait for more data; (c) complications of ASVAB restriction in range effects that are not simple to address, especially when selection ratios are stringent and the

criterion score distributions appear skewed; and (d) actual integrity of the criterion (which is not known until a full study is in progress). The best we might be able to do is to wait as long as possible within the practical parameters of the study for an adequate sample size to analyze, recognizing that prior validity studies for the same or similar occupations might provide some useful technical parameters.

Chapter 12. References

Borman, W. C., White, L. A., & Dorsey, D. W. (1995). Effects of ratee task performance and interpersonal factors on supervisor and peer performance ratings. *Journal of Applied Psychology, 80*, 168-177.

Borman, W. C., White, L. A., Pulakos, E. D., & Oppler, S. H. (1991). Models of supervisory job performance ratings. *Journal of Applied Psychology, 76*, 863-872.

Carretta, T. R., Perry, D. C., Jr., & Ree, M. J. (1996). Predicting situational awareness in F-15 pilots. *International Journal of Aviation Psychology, 6*, 21-41.

Cashen, L. H., & Geiger, S. W. (2004). Statistical power and the testing of null hypotheses: A review of contemporary management research and recommendations for future studies. *Organizational Research Methods, 7,* 151-157.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7,* 249-253.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (revised edition)*. Hillsdale, NJ: Erlbaum.

Cohen, J., & Cohen, P. (1983). *Applied multivariate regression/correlation analysis for the behavioral sciences (2nd edition)*. Hillsdale, NJ: Erlbaum.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. NY: Holt, Rinehart, and Winston.

Dyba, T., Kampenes, V., & Sjoberg, D. (2006). A systematic review of statistical power in software engineering experiments. *Information and Software Technology, 48,* 745-755.

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate behavioral research, 26,* 499-510.

Halpern S. D., Karlawish, J. H.T., & Berlin, J. A. (2002). The continuing unethical conduct of underpowered clinical trials. *Journal of American Medical Association, 288,* 358-62.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression (2nd edition)*. NY: Wiley.

Hunter, J. E. (1986). Cognitive ability, job knowledge, and job performance. *Journal of Vocational Behavior, 29,* 340-362.

Lance, C. E., & Bennett, W. (2000). Replication and extension of models of supervisory job performance ratings. *Human Performance, 13,* 139-158.

Lautenschlager , G. J., & Mendoza, J. L. (1986). A stepdown hierarchical multiple regression analysis for examining hypotheses about test bias in prediction. *Applied Psychological Measurement, 10,* 133-139.

Lawley, D. (1943). A note on Karl Pearson's selection formula. *Royal Society of Edinburgh, Proceedings, Section A, 62,* 28-30.

Lee, R., & Foley, P. P. (1986). Is the validity of a test constant throughout the test score range? *Journal of Applied Psychology, 71,* 641-644.

Mazen, A. M., Graf, L. A., Kellogg, C. E., & Hemmasi, M. (1997). Statistical power in contemporary management research. *Academy of Management Journal, 30,* 369-380.

Murphy, K. R., Myors, B., & Wolach, A. (2009). *Statistical power analysis* (3rd ed.). NY: Taylor and Francis Group.

Nunnally, J. (1978). *Psychometric theory* (2nd ed.). NY: McGraw-Hill.

Raju, N. S., Edwards, J. E., & LoVerde, M. A. (1985). Corrected formulas for computing sample sizes under indirect restriction in range. *Journal of Applied Psychological 70,* 565-566.

Raju, N. S., Steinhaus, S. D., Edwards, J. E., & DeLessio, J. (1991). A logistic regression model for personnel selection. *Applied Psychological Measurement, 15,* 139-152.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical regression: Applications and data analysis methods* (2nd ed.). Newbury Park: Sage.

Ree, M. J., Carretta, T. R., & Doub, T. W. (1998/1999). A test of three models of the role of g and prior job knowledge in the acquisition of subsequent job knowledge in training. *Training Research Journal, 4,* 135-150.

Ree, M. J., Carretta, T. R., & Teachout, M. S. (1995). Role of ability and prior job knowledge in complex training performance. *Journal of Applied Psychology, 80,* 721-730.

Sackett, P. R., & Wade, B. E. (1983). On the feasibility of criterion-related validity: The effects of range restriction assumptions on needed sample size. *Journal of Applied Psychology, 68,* 374-381.

Schmidt, F. L., Hunter, J. E., & Outerbridge, A. N. (1986). Impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. *Journal of Applied Psychology, 71,* 432-439.

Schmidt, F. L., Hunter, J. E., & Urry, V. W. (1976). Statistical power in criterion-related validation studies. *Journal of Applied Psychology, 61,* 473-485.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105,* 309-316.

Society for Industrial and Organizational Psychology (2003). *Principles for the validation and use of personnel selection procedures (4th edition)*. Bowling Green, OH: Author.

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology, 23,* 565-578.

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594-604.

Yaun, K-H., & Hayashi, K. (2003). Bootstrap approach to inference and power analysis based on three test statistics for covariance structure models. *British Journal of Mathematical and Statistical Psychology, 56,* 93-110.

Chapter 13.
Weighting Variables: The Tradeoff between Validity and Adverse Impact
Jeff W. Johnson

Introduction

With the exception of the Army, the Services use integer (unit) weights to construct their ASVAB occupational classification composites, which involve two to four tests (the Army uses full-least squares regression weights using all ASVAB tests). Several reasons have been proposed for using simple or unit weights, including simplifying computation (Stahlnaker, 1938), improving robustness (Dawes, 1979), reducing shrinkage under cross-validation compared to regression weights (Wainer, 1976, 1978), and providing better generalizability to future samples theoretically coming from the same population. This and the next chapter deal with issues to consider when weighting variables in prediction equations with this chapter focusing on a heuristic that organizations can use to assess the specific tradeoff of maximizing validity while minimizing adverse impact. We do not specifically address the methods used to assess adverse impact (a limited discussion is provided in Chapter 15) and so we refer the reader to the vast literature on the topic (e.g., Linn, 1973; Thorndike, 1971) some of which includes the military context (e.g., Wise et al., 1992).

To Weight or Not to Weight?

It is reasonable to ask if there is any point to differential weighting of tests in a composite. There is a large literature that suggests that unit or simple weights yield nearly the same results as regression weights (Aiken, 1966; Ree, Carretta, & Earles, 1998; Wainer, 1976, 1978; Wilks, 1938). Three factors are important for determining the expected correlation between composites of a set of tests – the average correlation among the tests, the number of tests, and the relative variability of the weights. Under typical circumstances, differently weighted combinations of the same tests into composites will yield results that are very highly correlated (Wilks, 1938).

For example, Ree et al. (1998) cited published examples of nearly identical rank orderings of individuals for composites that were based on regression weights, unit weights, policy-capturing weights, and factor weights. The more highly correlated the tests are, the more similar the rank-ordering of the individuals will be, even though the composites were formulated to weight the tests differentially (Ree et al., 1998). Because military enlistment and classification composites are based on the ASVAB, which consists of moderately intercorrelated tests, different weighting schemes are likely to make very little difference in the rank-ordering of individuals or the amount of criterion variance explained (i.e., R^2). If top-down selection is used, the same individuals will be selected regardless of the weights.

Prediction vs. Explanation

Multiple regression analysis has two distinct applications: prediction and explanation (Courville & Thompson, 2001). When multiple regression is used for a purely predictive purpose, the regression equation derived within a sample is used to predict scores on a criterion variable from a weighted combination of variables (i.e., test scores). This same equation can be applied later to test scores for a similar sample, or a future sample not yet available. The issue in the prediction application becomes whether the original equation is optimal in the new sample.

The elements of the equation are least-squares regression coefficients, which indicate the amount by which the criterion score is expected to change as the result of a unit increase in a given predictor score, while holding the other predictors constant. These regression coefficients minimize the sum of squared errors of prediction about the linear regression line and are optimal for maximizing prediction in the sample in which they were developed. Of greatest interest in the prediction application is the extent to which the criterion can be predicted by the predictor variables (indicated by R^2), with less interest in the relative magnitude of the regression coefficients.

When multiple regression is used for explanatory or theory-testing purposes, the interest is in the extent to which *each* variable contributes to the prediction of the criterion. For example, if theory suggested that one variable was relatively more important than another, we would expect this to be reflected in their relative regression weights. Interpretation of the regression weights is the primary concern, such that substantive conclusions can be drawn regarding one predictor with respect to another.

Least-squares regression coefficients are not designed to be interpreted in this way and are uninterpretable in terms of relative importance when predictor variables are correlated (Johnson & LeBreton, 2004). In this case, the appropriate procedure is dominance analysis (Budescu, 1993) or relative weight analysis (Johnson, 2000). These procedures partition the predictable variance in the criterion (represented by R^2) among the predictors according to the proportionate contribution each predictor makes, considering both its direct effect and its effect when combined with the other variables in the regression equation (Johnson & LeBreton).

Relative Weight Analysis

Johnson and LeBreton (2004) recommend two alternate methods for measuring the relative importance of predictors - Budescu's (1993) dominance analysis and Johnson's (2000) relative weight analysis (RWA). Both methods (a) yield importance weights that represent the proportionate contribution each predictor makes to R^2, (b) consider a predictor's direct effect and its effect when combined with other predictors, and (c) result in estimates of importance that make conceptual sense. The two methods also produce almost identical results despite using very different approaches to evaluate the importance of the predictors (Johnson, 2000; LeBreton, Ployhart, & Ladd, 2004).

The advantage of RWA over dominance analysis is that relative weights can be computed much more quickly than dominance analysis weights, both in terms of researcher time and computer processing time. RWA takes the same amount of time regardless of the number of predictors, but the time required to run a dominance analysis increases exponentially as the number of predictors increases. Dominance analysis requires that regression analyses be conducted for all possible combinations of predictors, so a 10-predictor model requires 1,023 separate regression analyses. Even with modern high-speed computers, this can take significant CPU time. Dominance analysis requires that code be written or edited, whereas RWA requires only syntax that specifies the included variables. Choosing RWA over dominance analysis could therefore result in considerable cost and time savings, especially when multiple analyses are required. For these reasons, empirical studies comparing relative importance weights to regression weights have used RWA. Thus, this chapter focuses on RWA.

RWA is based on the observation that most statistical measures of predictor importance yield the same results when predictors are uncorrelated. For example, squared zero-order correlations, squared standardized regression coefficients, Hoffman's (1960) product measure, and Darlington's (1968) usefulness are all equivalent and sum to R^2 when predictors are uncorrelated. Therefore, the first step in RWA is to transform the predictors (e.g., specific attributes measured on a survey) to their maximally related orthogonal counterparts. In other words, a set of new variables is created that are as highly related as possible to the original set of predictors but are uncorrelated with each other. Gibson (1962) describes this relatively simple RWA mathematical process.

Conceptually, the RWA process could be likened to a principal components analysis in which the same number of components as number of predictors is extracted and rotated to the point where no other rotation would yield higher correlations between each original predictor and its associated orthogonal variable. The criterion (i.e., some overall evaluation measured by the survey such as overall customer satisfaction or overall employee satisfaction) is then regressed on the new uncorrelated variables. The squared standardized regression coefficients unambiguously represent the relative importance of the new variables.

The relative importance of the *new* variables is an approximation of the relative importance of the *original* predictors. To arrive at an estimate of the relative importance of the *original* predictors, there must be some mechanism by which information on the relations between the *new* variables and the *criterion* is combined with information on the relations between the *original* predictors and the *new* variables. Johnson (2000) showed that the appropriate way to do this was to regress the original predictors on the new orthogonal variables. Because regression coefficients are assigned to the uncorrelated variables, the relative importance of the uncorrelated variables to the original predictors is unambiguous.

By combining the indices representing the relative importance of the uncorrelated variables to the criterion and the indices representing the relative importance of the uncorrelated variables to the original predictors, we can compute an index representing the relative importance of the original predictors to the criterion (i.e., relative weights). In practical terms, the output of RWA is a weight for each predictor that represents its relative contribution to the dependent variable. Larger weights indicate a stronger association with the outcome.[11]

Using Relative Weights for Prediction

Least-squares regression coefficients that maximize the prediction of a criterion have long been considered inadequate as measures of predictor importance, especially under conditions of multicollinearity (Budescu, 1993; Green & Tull, 1975; Hoffman, 1960). Conversely, there is some question as to whether relative importance indices are appropriate for use as weights in situations where the primary concern is prediction. Least-squares regression coefficients yield the highest possible R^2 in the sample in which they were derived. The question of ultimate interest, however, is how well these regression coefficients predict when they are applied to data in another sample. Two primary factors (sampling error and multicollinearity) influence the extent to which regression coefficients derived in one sample will predict in another. Both suggest that relative importance weights may have an advantage over least-squares regression weights in some circumstances.

Sampling error makes least-squares regression weights prone to inaccuracy, especially when sample sizes are very small. Therefore, unit weights (i.e., all predictors weighted equally) are often superior when applied to population data (Schmidt, 1971). In other words, regression weights in a small sample could be so different from the optimal weights in the population that it is better not to weight the predictors at all. Even when sample sizes are large, unit weights are often applied in practice because organizational stakeholders perceive them as easy to explain and maintain over time. The simplicity and interpretability of unit weights very often trump the better prediction of regression weights.

High multicollinearity (i.e., intercorrelations between predictor variables) leads to instability in regression weights, making them less applicable outside the sample in which they were derived (Wainer, 1978). Dominance analysis and RWA are designed to provide estimates of relative importance precisely under conditions of multicollinearity, so they are more likely to be stable across samples. Greater stability of relative importance weights under conditions of multicollinearity makes it reasonable to hypothesize that sample-based relative importance weights may sometimes provide better prediction in the population than do sample-based least-squares regression weights. This is especially likely in small samples where regression weights are dependent on the idiosyncrasies of the sample data.

[11] See Johnson (2000) for mathematical formulas detailing the derivation and calculation of relative weights.

In addition to the statistical rationales for using relative importance weights for prediction over regression or unit weights, there are also conceptual considerations that give importance weights an advantage. Relative importance weights can be expressed as the percentage of predictable criterion variance attributed to each predictor, so these indices may be easier to present to decision makers (e.g., managers, executives, board members) when compared to regression coefficients or increments in R^2 (LeBreton, Hargis, Griepentrog, Oswald, & Ployhart, 2007). For example, it is likely much easier to convince organizational decision makers to invest money in a new selection instrument when it is described as accounting for 25% of the predictable variance in overall job performance than when it is described as increasing R^2 by .03. On the other hand, as was demonstrated in Chapter 3, an R^2 translates into a validity coefficient. The increase in R^2 of .03 translates into an increment in validity of .17 which, with much explanation, can be used to project expected improvments in, say, military training success rates and the cost savings by not having to (a) re-recruit individuals to fill the slots of those failed (b) reassign failed students to another occupation, (c) transport them to other training sites, and (d) realize a much-diminshed time in productive status (during their first term of enlistment).

Empirical Studies

At least two studies have compared the predictive power of relative importance indices to least-squares regression coefficients. Oswald, Johnson, and Oliver (2000) compared relative weights (Johnson, 2000) to regression weights, unit weights, and rational weights (rational weights were estimates of the relative importance of predictors made by 26 industrial-organizational psychologists). Using correlation matrices between 9 predictors and each of 11 criteria as the population correlation matrices, these authors conducted a Monte Carlo study in which least-squares regression weights and relative weights were computed within 1,000 replications of each of four sample sizes (n = 50, 100, 200, or 500). These weights were then applied to the population matrix, and R^2 was computed in each instance.

Oswald et al. (2000) also computed R^2 when applying unit weights and rational weights. Results indicated that, on average, (a) relative weights were less biased than least-squares regression weights, (b) relative weights were more stable than least-squares regression weights, (c) relative weights tended to be more similar to rational weights than were least-squares regression weights, and (d) sample-based relative weights tended to yield a higher R^2 in the population than did unit weights and rational weights. Sample-based relative weights also tended to yield a higher R^2 in the population than did least-squares regression weights with smaller sample sizes (less than about 100).

Oswald (2001) followed up this study with another Monte Carlo study in which the number of predictors, extent of multicollinearity among the predictors, and sample size were varied. Further, range restriction and criterion and predictor reliability artifacts were built into some of the conditions. Oswald found that relative weights tended to provide better prediction than regression weights for sample sizes of 100 or less when multicollinearity was relatively low but not when multicollinearity was high. The

superiority of regression weights under high multicollinearity conditions was explained by the presence of suppressor variables. Suppressor variables have low zero-order correlations with the criterion but large negative regression coefficients because they suppress error variance in the other predictors, thereby improving prediction. Suppressor variables are common when all predictors are highly intercorrelated. Multiple regression analysis is designed to take advantage of suppressor variables to enhance prediction, but relative weight analysis is not. Multiple regression had less of an advantage when range restriction and unreliability artifacts were added. When there were few predictors and sample sizes were small, relative weights demonstrated better prediction despite the presence of suppressor variables.

Adverse Impact of a Composite

A common problem faced by personnel selection researchers and practitioners involves choosing a set of predictors from a larger set of potential predictors for the purpose of creating a selection test battery. There are usually two primary considerations when creating a test battery: maximizing the criterion-related validity of the test battery while minimizing adverse impact against protected groups. By adverse impact we refer to majority and minority mean differences in test scores on a selection instrument favoring the majority group such that the majority group is hired at a higher rate. (Adverse impact is discussed in the Chapter 18). Creating a composite of several valid predictors is a common strategy for reducing the degree to which a selection procedure produces group differences (Campbell, 1996; Sackett & Ellingson, 1997; Schmitt, Rogers, Chan, Sheppard, & Jennings, 1997). The problem is that the most valid predictors of performance tend to also produce the largest group differences (Sackett & Ellingson, 1997), so adding a predictor that increases the validity of the composite will often have the simultaneous effect of increasing adverse impact.

Compounding the problem is the fact that adding a predictor with little adverse impact to a predictor with large adverse impact typically does not reduce the adverse impact of the composite to the extent that generally would be expected (Sackett & Ellingson, 1997). Sackett and Ellingson provided an example of two uncorrelated predictors. One predictor had a standardized mean subgroup difference (*d*) of 1.00 and the other had a *d* of 0.00. Most researchers would expect that the two predictors would offset each other, so the *d* of an equally weighted composite of the two predictors would be 0.50. In fact, the *d* of this composite would be 0.71. The composite *d* approaches the mean of the individual test *d*'s as the correlation between the tests increases. Thus, reducing adverse impact by adding predictors to a composite is not as easy as it seems at first glance.

Considering Both Validity and Adverse Impact

When stakeholders value both maximizing validity and minimizing adverse impact, the dilemma of the researcher evaluating alternative composites is determining at what point the gain in validity is offset by the increase in adverse impact. The problem is

exacerbated when there are many predictors from which to choose. The number of possible composites that can be created from p predictors is $2^p - 1$.[12] With just 5 predictors, there are 31 potential composites to evaluate. It is difficult to compare 31 composites to determine which is best at balancing validity and adverse impact. With just a few more predictors, the number of potential composites increases rapidly. With 8 predictors, the number of potential composites is 255. With 10 predictors, it increases to 1,023 (we are reminded that the ASVAB has 9 tests at this point). Although some composites can be rejected immediately because they are obviously inferior, there usually will be a large number of composites among which it could be difficult to choose.

It would be desirable to have an automated procedure for choosing among composites that considers both validity and adverse impact. This would make the process of choosing a composite much less complicated, take much of the subjectivity out of the process, and make the selection procedure easier to defend. Johnson, Abrahams, and Held (2004; see also Johnson & Abrahams, 2003) proposed a procedure to select predictors for composites that considers both criterion-related validity and standardized mean subgroup differences. This procedure was flexible enough to allow the user to adjust the parameters depending on the relative value placed on validity and adverse impact.

Johnson et al. (2004) recognized that a larger increase in validity should be required to justify increasing adverse impact when adverse impact is already relatively high than when it is low. For example, an increase in the standardized mean subgroup difference (d) from 0.00 to 0.10 is not as damaging as an increase from 0.50 to 0.60. The increases are of the same magnitude, but in the first case, adverse impact is still low and in the second case it is becoming less acceptable. A small increase in validity is worthwhile if adverse impact is low, but a larger increase in validity should be required if it is already high. The solution was to create a formula that could be applied to d that would transform it to a value that decreases exponentially as d increases. Less adverse impact results in a higher transformed score. By adding the validity coefficient to this transformed d value, a choice can be made between alternative composites by choosing the one that has the highest combined validity/transformed adverse impact sum. The idea behind the transformed d score is that, as adverse impact increases, it takes a progressively larger increase in validity to justify choosing the composite.

After experimenting with several transformation formulas, Johnson et al. (2004) determined that the following formula best represented their conceptualization of the relative value of validity and adverse impact at different levels of d:

$$d_t = 1 - \left(\frac{ad + bd^2}{25} \right). \quad (13\text{-}1)$$

[12] Note that p of these "composites" will comprise just the single tests (i.e., they will involve just one measure).

Parameters a and b are similar to the constant and slope in a regression equation. Parameter a, like the constant, affects the starting point for the increase in validity required to offset a given increase in adverse impact. Parameter b, like the slope, affects the rate at which the required increase in validity increases as d increases. Users of this procedure can adjust these parameters to reflect their personal or institutional preferences for the relative weight that should be placed on adverse impact and validity when choosing test composites.

As an example of the use of Equation 13-1, set $a = 4$ and $b = 5$. Suppose adding a subtest increases the adverse impact of the composite from 0.10 to 0.20. For the sum of d_t and validity to increase beyond the sum for the previous composite, validity would have to increase by at least .022 – that is, the difference in d_t values calculated at $d = .10$ (.982) and $d = .20$ (.960). Similarly, if adverse impact increased from 0.50 to 0.60, validity would have to increase by at least .038. This takes into account the fact that increases in adverse impact when there is very little adverse impact are more acceptable than increases in adverse impact when adverse impact is more severe. To determine the combined validity/adverse impact score (*VAI*) for a particular composite, Johnson et al. (2004) used the following equation:

$$VAI = r_c + d_t, \quad (13\text{-}2)$$

where r_c is the validity of the composite and d_t is as defined in Equation 13-1. The best composite is the one in which *VAI* is at its maximum. Johnson et al. noted that d_t can be replaced by the mean d_t across different subgroup comparisons (e.g., Black-White, Hispanic-White, Male-Female) if more than one comparison is of concern. A weighted mean d_t also can be computed if certain types of adverse impact are of greater concern than are others.

Although a great deal more research is necessary in this area, the Johnson et al. (2004) procedure has some promise as a starting point in automating the selection of composites when considering both validity and adverse impact. This procedure has been used operationally in a study in which one of many possible predictor composites had to be chosen for each of several criterion variables (Johnson, Paullin, & Hennen, 2005). In this study, VAI was computed for each possible composite for each criterion and was one piece of information used in choosing the final composite for each criterion. Future research should be directed toward further refining the VAI formula and identifying more objective ways of placing relative value on validity and adverse impact.

Concluding Remarks

Given the mixed results regarding weights, it is not clear under what conditions relative weights provide better prediction than least-squares regression weights. Given the distinct advantages that relative importance indices have over regression coefficients in terms of communicating to decision makers, it might be worth exploring the possibility that relative weights have predictive advantages. Also, considering both validity and adverse impact simultaneously is feasible but involves policy makers and

technical information about minority group score barriers that result from the current set of selection instruments. The other strategy for reducing score barriers, as adopted by the Navy, is to offer alternative ASVAB composites that, with appropriate cutscores, form standards that apply to everyone. This way the Navy can capitalize on the prior experience/knowledge in the technical areas that some recruits have, but maintain a classification system that maintains diversity across most Ratings. The next chapter focuses on the calculation of weights with a focus on a composite of criteria that can be used in test validation research.

Chapter 13. References

Aiken, L. R., Jr. (1966). Another look at weighting test items. *Journal of Educational Measurement, 3, 183-185.*

Budescu, D. V. (1993). Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin*, *114*, 542-551.

Campbell, J. P. (1996). Group differences and personnel decisions: Validity, fairness, and affirmative action. *Journal of Vocational Behavior, 49,* 122-158.

Courville, T., & Thompson, B. (2001). Use of structure coefficients in published multiple regression articles: β is not enough. *Educational and Psychological Measurement*, *61*, 229-248.

Darlington, R. B. (1968). Multiple regression in psychological research and practice. *Psychological Bulletin*, *69*, 161-182.

Dawes, R. (1979). The robust beauty of improper linear models. *American Psychologist, 34,* 571-582.

Gibson, W. A. (1962). Orthogonal predictors: A possible resolution of the Hoffman-Ward controversy. *Psychological Reports*, *11*, 32-34.

Green, P. E., & Tull, D. S. (1975). *Research for marketing decisions* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Hoffman, P. J. (1960). The paramorphic representation of clinical judgment. *Psychological Bulletin*, *57*, 116-131.

Johnson, J. W. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate Behavioral Research*, *35*, 1-19.

Johnson, J. W., & Abrahams, N. (2003). *Exploring alternative methods of creating and weighting ASVAB composite component tests for classifying personnel into U.S. Navy jobs* (Institute Report #434). Minneapolis, MN: Personnel Decisions Research Institutes, Inc.

Johnson, J. W., Abrahams, N., & Held, J. D. (2004, April). *A procedure for selecting predictors considering validity and adverse impact*. Poster presented at the 19th Annual Conference of the Society for Industrial and Organizational Psychology, Chicago.

Johnson, J. W., & LeBreton, J. M. (2004). History and use of relative importance indices in organizational research. *Organizational Research Methods*, *7*, 238-257.

Johnson, J. W., Paullin, C., & Hennen, M. (2005). Validation and development of an operational version of Exam 473. In C. Paullin & J. W. Johnson (Eds.), *Development and validation of Exam 473 for the United States Postal Service* (Institute Report #496). Minneapolis, MN: Personnel Decisions Research Institutes, Inc.

LeBreton, J. M., Hargis, M. B., Griepentrog, B., Oswald, F. L., & Ployhart, R. E. (2007). A multidimensional approach for evaluating variables in organizational research and practice. *Personnel Psychology*, *60*, 475-498.

LeBreton, J. M., Ployhart, R. E., & Ladd, R. T. (2004). A Monte Carlo comparison of relative importance methodologies. *Organizational Research Methods*, *7*, 258-282.

Linn, R. L. (1973). Fair test use in selection. *Review of Educational Research, 43,* 139-161.

Oswald, F. L. (2001, April). Weighting tables: Results comparing different multiple regression weighting methods. In J. M. LeBreton & J. W. Johnson (Chairs), *Use of relative importance methodologies in organizational research*. Symposium conducted at the 16th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Oswald, F. L., Johnson, J. W., & Oliver, D. H. (2000, April). *The importance of relative importance weights: Statistical and rational considerations*. In J. W. Johnson (Chair), Practical applications of relative importance methodology in I/O psychology. Symposium conducted at the 15th Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Ree, M. J., Carretta, T. R., & Earles, J. A. (1998). In top-down decisions, weighting variables does not matter: A consequence of Wilks' theorem. *Organizational Research Methods, 1,* 407-420.

Sackett, P. R., & Ellingson, J. E. (1997). The effect of forming multi-predictor composites on group differences and adverse impact. *Personnel Psychology*, *50*, 707-721.

Schmidt, F. L. (1971). The relative efficiency of regression and simple unit predictor weights in applied differential psychology. *Educational & Psychological Measurement, 31,* 699-714.

Schmitt, N., Rogers, W., Chan, D., Sheppard, L., & Jennings, D. (1997). Adverse impact and predictive efficiency of various predictor combinations. *Journal of Applied Psychology, 82,* 719-730.

Stahlnacker, J. (1938). Weighting questions in the essay-type examination. *Journal of Educational Psychology, 7,* 481-490.

Thorndike, R. L. (1971). Concepts of culture-fairness. *Journal of Educational Measurement, 3,* 63-70.

Wainer, H. (1978). On the sensitivity of regression and regressors. *Psychological Bulletin*, *85*, 267-273.

Wilks, S. S. (1938). Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika, 3,* 23-40.

Wise, L., Welsh, J., Grafton, F., Foley, P., Earles, J., Sawin, L., & Divgi, D. R. (1992). *Sensitivity and fairness of the Armed Services Vocational Aptitude Battery (ASVAB) technical composite (*DMDC Technical Report 92-002). Seaside, CA: Defense Manpower Data Center.

Chapter 14.
More on Weights: Forming a Composite of Multiple Performance Criteria
Rodney A. McCloy

Introduction

This chapter extends the discussion on weights in the last chapter with a slightly different emphasis and computational framework. Methods of determining weights for a set of components that are to be combined to form a composite (whether a predictor composite or a criterion composite) can be thought of as falling into two primary categories – rational and empirical weights. Rational weights are developed by judges who provide numeric weights based on their views about how the components should contribute to the composite (often based on each judge's notions of what organizational policy is or should be). Empirical weights are developed when quantitative techniques are applied to yield a set of weights that achieve some pre-determined goal (e.g., maximize the reliability of the resulting composite). Sometimes, empirical weights are obtained after performing mathematical operations on data provided by judges/stakeholders, thus representing both categories to a degree. This chapter provides a discussion of two rational weighting approaches and one empirical method that uses judgments as data. It concludes by identifying additional weighting goals that might lead one to compute empirical weights. Also, the shift is from training performance as the criterion, which has been the focus of the Navy's approach in ASVAB validation/standards studies, to the dimensions of job performance as the "criteria."

Rational Weights: Direct Estimation

When creating a composite variable (assume for now we are creating a composite of performance measures), component weights often can be provided with relatively little muss and fuss. An individual or group of policy makers can simply agree that all the components in the performance composite should be weighted equally. If so, the components are said to be "unit-weighted" (as discussed in the previous chapter). A unit-weighting scheme most often arises when there is an explicit desire to treat all the components as equal to one another in importance, although it can sometimes arise when there is no clear policy for how best to assign weights to the components.

Another common approach is to assemble a group of expert judges and have them allocate 100 percentage points across the components. Sometimes judges will make the point assignments individually and then review the results as a group before arriving at a final set of weights, which is most often the mean across judges (sometimes a second round of individual weighting is conducted following the group session). Unit weights can be viewed as a special case of percentage allocation, with all the components receiving an equal percentage contribution to the composite (i.e., each weight is equivalent to $100/k$, where k is the number of components to be weighted).

Whatever the weights, the purpose of these *a priori* weighting schemes (Wang & Stanley, 1970) is to reflect the policies of the judges participating in the weighting exercise. Perhaps the most important point regarding weights obtained via direct estimation is that the weights allow the composite to be defined as desired. That is, the weights are not driven by statistical realities or limitations, but rather by policy makers and other stakeholders who have the opportunity to define the reality they envision (by creating a composite score that reflects their overall view).

Rational Weights: Policy Specifying

Another means of developing rational weights involves a procedure known as "policy specifying" (Ward, 1977). Policy specification is "a decision-modeling technique by which variables identified as pertinent to a decision-making process can be combined to derive a single predicted payoff value" (Piña, Emerson, Leighton, & Cummings, 1988, p.5). The U.S. Air Force used this procedure to weight and combine 10 variables deemed critical to the assignment of enlisted personnel to Air Force Specialties into a single score (payoff function) that the Air Force then used with their Processing and Classification of Enlistees (PACE; Piña et al., 1988) system used to quantify the efficacy of the Air Force's person-job matching decisions.

Policy specification is a form of "policy capturing", a general term that describes methods for obtaining policy information from stakeholders and decision-makers, and describing the relations between that information and the judgments based on it (Rogelberg, Ployhart, Balzer, & Yonker, 1999). With regard to PACE, the procedure began with the assembly of a group of subject matter experts (classification experts, policy makers) who held weekly meetings during which they discussed and eventually identified the variables they deemed most important to making sound assignment decisions. The variables included enlistee aptitude, training preferences, gender, training cost, probability of completing the first term of enlistment, and the fill priorities of the available jobs.

Having identified 10 such variables and selected measures of them, the Air Force judges then combined these measures into groups in a bottom-up agglomerative process. For example, they formed a trainability score that comprised intellectual ability and academic background. The aggregation continued until all 10 variables had been combined into a single score that served as the index upon which the value of various classification decisions were based. The grouped variables were weighted by importance at each step.

Piña et al. (1988) did not describe how the importance weights were determined. Regardless, the importance weights and the groups of variables formed by the expert panel serve as rationally-weighted representations of the policies of the panel. The policy-specifying technique could serve as a basis for judgmentally determining the underlying value stakeholders ascribe to multiple criteria that need to be combined to form a single composite criterion.

Empirical Weights: Conjoint Measurement

Another policy-capturing approach develops empirical weights for the components; however, these empirical weights are derived from expert judgments about the relative importance/worth of the components. Conjoint scaling (Green & Srinivasan, 1978; Johnson, 1974) (sometimes called conjoint measurement) requires judges to evaluate (e.g., rank-order, rate) groups of stimuli that differ systematically with regard to the dimensions in question. The stimuli may differ on all the dimensions of interest simultaneously (full-profile conjoint scaling) or on just two dimensions (two-factors-at-a-time conjoint scaling).

As an example of the latter approach to conjoint scaling, Sadacca, Campbell, White, and DiFazio (1989) sought to determine the relative importance of weights that should be assigned to the five job performance constructs developed during the Army's Project A (Campbell, 1990; Campbell & Knapp, 2001) to yield a composite representing overall job performance. The five constructs of interest were Core Technical Proficiency, General Soldiering Proficiency, Effort and Leadership, Maintaining Personal Discipline, and Physical Fitness and Military Bearing (Campbell, 1986).

In the Sadacca et al. (1989) study, non-commissioned officers and company-grade and field-grade officers representing 20 Army occupations served as judges. The judges' task was to rank 15 hypothetical Soldiers in terms of overall job performance. The judges provided rankings for 10 sets of 15 Soldiers. The 15 Soldiers within each set differed in terms of their relative standing on two of the Project A performance constructs (e.g., Effort and Leadership vs. General Soldiering Proficiency). If Soldiers who scored higher on a given performance construct (say, Effort and Leadership) than another (say, General Soldiering Proficiency) were ranked higher than Soldiers who had the opposite scoring pattern, then the construct with the higher score among the more highly ranked Soldiers was deemed more important (and thus received a larger empirical weight) than the other construct.

Conjoint scaling produces weights that are based on the ratio of the dimension (here, construct) regression weights obtained when predicting a given judge's ranking of the stimuli (here, Soldiers) (see Torgerson, 1958). Sadacca et al. (1989) found that the weights applied to the five performance components differed significantly across the 20 jobs (consistent with differential assignment theory, which is discussed in the last chapter). Conjoint scaling could assist the Navy and other Services with determining a reasonable set of weights to apply to multiple components of a composite performance measure.

Empirical Weights: Other Goals

As with selection and classification goals, there may be many goals to consider in weighting components in a performance composite. Wang and Stanley (1970) present many options for weighting components in a composite, including weighting (a) to achieve equal correlations of the components with the resulting composite, (b) to achieve minimum variation of the composite, (c) to achieve maximum reliability of the

resulting composite, (d) by difficulty (of test/measure), and (e) by length (of test/measure). The choice of the appropriate weighting scheme depends on the purpose of the composite itself—in particular, whether it is to be viewed as a psychological construct or else as a statistical conglomerate that might embody multiple dimensions for the purpose of encompassing as many relevant factors as possible when yielding a score that will serve as a decision index.

What You See Is (Probably) Not What You Get: Nominal Weights and Effective Weights

Frequently, the assignment of a set of desired weights to multiple components (whether predictors or criteria) does not yield the intended weighting scheme. Rather, these *nominal* weights specified by the weighting scheme (e.g., equal weights, differential weights assigned by expert judges) will contribute to, but not equal (and likely will not even be proportional to), the *effective* weights that result when the composite is formed (Wang & Stanley, 1970). That is, the nominal weights transform into a different set of weights. The transformation is a function of the variances of and covariances among the elements constituting the composite. This occurs because the variance of a composite is a function of the sum of the variances of the components and their covariances, as shown in Equation 1 (p. 664) from Wang and Stanley:

$$Var(X_1 + X_2 + \ldots + X_n) = Var(X_1) + Var(X_2) + \ldots + Var(X_n) + 2Cov(X_1X_2) + 2Cov(X_1X_3) + \ldots + 2Cov(X_{n-1}X_n).$$

When nominal weights (w) have been assigned to the various components, the variance of the weighted composite is thus

$$Var(w_1X_1 + w_2X_2 + \ldots + w_nX_n) = w_1^2Var(X^1) + w_2^2Var(X_2) + \ldots + w_n^2Var(X_n) + 2w_1w_2Cov(X_1X_2) + 2w_1w_3Cov(X_1X_3) + \ldots + 2w_{n-1}w_nCov(X_{n-1}X_n).$$

(Wang & Stanley, Equation 3, p. 665). This, in turn, means that the contribution of any single component from that composite to the variance of the composite is

$$C_i = w_iw_1Cov(X_iX_1) + w_iw_2Cov(X_iX_2) + \ldots + w_iw_{i-1}Cov(X_iX_{i-1}) + w_i^2Var(X_i) + w_iw_{i+1}Cov(X_iX_{i+1}) + \ldots + w_iw_nCov(X_iX_n).$$

(Wang & Stanley, Equation 4, p. 665). Thus, although the nominal weights certainly contribute to the resulting effective weights, they do not equal them unless the components are uncorrelated or correlate with one another to the same degree. Further, the formula indicates that as components are added, the variance of the composite is increasingly a function of the covariances rather than of the variances.

Wang and Stanley (1970) pointed out that this discrepancy between nominal and effective weights will occur whenever normative scoring is in effect—that is, when the meaning of a given test score is interpreted relative to scores from others on the test (e.g., percentiles, standard scores). If a criterion-referenced scoring process is used (e.g., percent correct), then effective weights typically will be proportional to nominal weights.

Fortunately, given a set of nominal weights, one can solve for values for the empirical weights that, when applied to the variances and covariances, will yield effective weights for the components that equal the desired, intended nominal weights. To demonstrate, assume we have scores on five performance dimensions with the following variance-covariance (VCV) and correlation (Corr) matrices (Tables 14-1 and 14-2, respectively):

Table 14-1
Variance/Covariance (VCV) Matrix for Five Performance Dimensions

	1	2	3	4	5
1	32				
2	20	27			
3	12	4	41		
4	16	6	22	36	
5	6	10	7	13	15

Table 14-2
Correlation (Corr) Matrix for Five Performance Dimensions

	1	2	3	4	5
1	1				
2	.68041	1			
3	.33129	.12022	1		
4	.47140	.19245	.57264	1	
5	.27386	.49690	.28227	.55943	1

Also assume that we would like these five dimensions to receive equal weight when forming a performance composite. The task of obtaining empirical weights becomes one of solving a system of five equations for five unknowns that, when applied to the variances and covariances as specified in the equations that follow, will yield effective weights that are equivalent to the desired magnitude expressed by the nominal weights. The five equations are as follows:

$$X1: 1 = (w1)^2*Var(X1) + (w1*w2)*Cov(X1,X2) + (w1*w3)* Cov(X1,X3) + (w1*w4)*Cov(X1,X4) + (w1*w5)*Cov(X1,X5)$$

$$X2: 1 = (w2)^2*Var(X2) + (w2*w1)*Cov(X2,X1) + (w2*w3)* Cov(X2,X3) + (w2*w4)*Cov(X2,X4) + (w2*w5)*Cov(X2,X5)$$

$$X3: 1 = (w3)^2*Var(X3) + (w3*w1)*Cov(X3,X1) + (w3*w2)* Cov(X3,X2) + (w3*w4)*Cov(X3,X4) + (w3*w5)*Cov(X3,X5)$$

$$X4: 1 = (w4)^2*Var(X4) + (w4*w1)*Cov(X4,X1) + (w4*w2)* Cov(X4,X2) + (w4*w3)*Cov(X4,X3) + (w4*w5)*Cov(X4,X5)$$

$$X5: 1 = (w5)^2*Var(X5) + (w5*w1)*Cov(X5,X1) + (w5*w2)* Cov(X5,X2) + (w5*w3)*Cov(X5,X3) + (w5*w4)*Cov(X5,X4)$$

where *X1* through *X5* are the five component scores to be weighted and *w1* through *w5* are the empirically determined weights for which we seek a solution.

When not constrained to equal 1.0 (as is specified for these five equations), the formula on the right side of the equals sign in these equations yields the effective weight for the particular component under consideration, given application of the nominal weights (here, $w1 = w2 = w3 = w4 = w5 = 1.0$) to the variances and covariances (and is equivalent to Wang and Stanley's [1970] Equation 4 presented earlier). In the present example, each component's sum is simply the sum of the variance and covariances for a given component (because all weights equal 1.0). Thus, for component *X1*, applying weights of 1.0 yields an effective weight of 86, which is the sum of 32 (its variance) and 20, 12, 16, and 6 (its covariances with the other four components).

Table 14-3 presents the effective weights for all five components in this example, along with the nominal (desired) weight for each component ("Nominal Weight") and the percentage contribution to the composite that the nominal weights specify each component should have ("Nominal %").

Table 14-3
Nominal, Effective, and Empirical Weights for Five Performance Components: Equal Nominal Weights and Unstandardized Variables

Component	Nominal Weight	Nominal %	Effective Weight	Effective %	Empirical Weights
X1	1	20	86	22.5	0.104
X2	1	20	67	17.5	0.124
X3	1	20	86	22.5	0.107
X4	1	20	93	24.3	0.097
X5	1	20	51	13.3	0.160

In Table 14-3, dividing each component's effective weight by the sum of all of the effective weights yields the percentage contribution of each component to the composite (the "Effective %" column in Table 14-3). These percentages clearly do not equal the desired apportionment of 20% (1/5) to all five components.

Because of the differing variances and covariances among the components, the weights that need to be assigned to the components to obtain equal effective weights would *not* be 1.0 across the board (as the nominal weights might suggest). Rather, the solution for this system of equations given the variance-covariance matrix (Table 14-1) is listed in the "Empirical Weights" column of Table 14-3.[13] Applying these empirically determined weights to the components *X1* through *X5* will yield effective weights for the components that match the desired nominal weights of $w1 = w2 = w3 = w4 = w5 = 1.0$. Thus, applying the empirical weights to the variances and covariances will result in each component contributing equally to the composite.

A similar analysis can be achieved if the nominal weights specify that some components should contribute more to the composite than others. Assume that an expert judgment exercise yielded the following set of weights for the five components:

$$w1 = 20\%, w2 = 10\%, w3 = 10\%, w4 = 40\%, w5 = 20\%.$$

With this weight specification, the equations to solve would be as follows:

[13] The weights were obtained using the "Solver" add-in from Excel 2007. Other linear programming software also could calculate the weights that solve the specified system of equations.

*X1: 2 = $(w1)^2$*Var(X1) + (w1*w2)*Cov(X1,X2) + (w1*w3)* Cov(X1,X3) + (w1*w4)*Cov(X1,X4) + (w1*w5)*Cov(X1,X5)*

*X2: 1 = $(w2)^2$*Var(X2) + (w2*w1)*Cov(X2,X1) + (w2*w3)* Cov(X2,X3) + (w2*w4)*Cov(X2,X4) + (w2*w5)*Cov(X2,X5)*

*X3: 1 = $(w3)^2$*Var(X3) + (w3*w1)*Cov(X3,X1) + (w3*w2)* Cov(X3,X2) + (w3*w4)*Cov(X3,X4) + (w3*w5)*Cov(X3,X5)*

*X4: 4 = $(w4)^2$*Var(X4) + (w4*w1)*Cov(X4,X1) + (w4*w2)* Cov(X4,X2) + (w4*w3)*Cov(X4,X3) + (w4*w5)*Cov(X4,X5)*

*X5: 2 = $(w5)^2$*Var(X5) + (w5*w1)*Cov(X5,X1) + (w5*w2)* Cov(X5,X2) + (w5*w3)*Cov(X5,X3) + (w5*w4)*Cov(X5,X4)*

(note how the nominal weights serve as constraints on the left side of the equals sign). The effective weights for these specified nominal weights again are not as desired. Once more (a) applying the desired weights to the right side of each equation, (b) calculating the sum for each component, and then (c) dividing that sum by the total of the sums of all five components, we see the extent to which the application of the nominal weights to the five components "as they are" yields effective weights that differ from what is desired, shown in Table 14-4.

Table 14-4
Nominal, Effective, and Empirical Weights for Five Performance Components: Unequal Nominal Weights and Unstandardized Variables

Component	Nominal Weight	Nominal %	Effective Weight	Effective %	Empirical Weights
X1	2	20	344	19.4	0.153
X2	1	10	115	6.5	0.102
X3	1	10	171	9.7	0.080
X4	4	40	920	51.9	0.243
X5	2	20	222	12.5	0.222

As before, Table 14-4 shows that there is a departure from the desired apportionment of components to the composite. To recapture the desired nominal weights, we need to apply empirically determined weights (Footnote 13) to the variances and covariances of the components.

The problem regarding differences between desired nominal weights (e.g., those specified by expert judges) and effective nominal weights (those that actually exist because of the influence of the components' variances and covariances) does not vanish if one standardizes the components prior to weighting. Standardization will obviate the complicating problem of unequal variances across components, but unequal covariances will almost certainly remain, thus providing a similar (if somewhat lesser) problem.

For example, assume that we treated the correlation matrix (Table 14-2) as the variance-covariance matrix. As such, it means that we are using standardized variables. The problem of effective weights that stray from the desired specified weights remains, although the differences between desired and actual contributions of components to the composite are not as great as in the unstandardized case (Table 14-3). The standardized case with equal nominal weights is presented in Table 14-5.

Table 14-5
Nominal, Effective, and Empirical Weights for Five Performance Components: Equal Nominal Weights and Standardized Variables

Component	Nominal Weight	Nominal %	Effective Weight	Effective %	Empirical Weights
X1	1	20	2.76	21.3	0.589
X2	1	20	2.49	19.2	0.646
X3	1	20	2.31	17.8	0.683
X4	1	20	2.80	21.6	0.581
X5	1	20	2.61	20.2	0.618

Although standardizing the components prior to weighting reduced the discrepancy between nominal and effective weights (see Tables 14-5 and 14-6), as noted, it still did not produce the equal weighting we were seeking to achieve. As Wang and Stanley (1970) pointed out, "using nominal weights with standard scores probably comes closest to achieving equal effective weighting, particularly if the average correlation of each variable with the others is nearly constant" (p. 666). The empirically determined weights are still necessary to faithfully reproduce the desired weighting of the components.

Finally, to complete the comparison, Table 14-6 presents the results that would be obtained by applying the expert judges' nominal weights to standardized components.

Table 14-6
Nominal, Effective, and Empirical Weights for Five Performance Components: Unequal Nominal Weights and Standardized Variables

Component	Nominal Weight	Nominal %	Effective Weight	Effective %	Empirical Weights
X1	2	20	10.89	18.7	0.863
X2	1	10	4.24	7.3	0.530
X3	1	10	4.64	8.0	0.515
X4	4	40	27.10	46.9	1.458
X5	2	20	11.13	19.1	0.861

Concluding Remarks

The weights required to yield desired nominal weights are almost never the nominal weights. Achieving a desired allocation of influence of components on a composite therefore requires more than development of nominal weights. The nominal weights must be appropriately transformed into new weights that, along with the variances and covariances of the components, yield the desired nominal weighting for those components.

Below are a couple of summary points to keep in mind when creating a composite variable—whether a criterion composite (as discussed in this chapter using job performance as the example) or a predictor composite. Following these suggestions will enhance the interpretability of the composite and ensure you are using a variable that has the properties you desire it to have.

- If forming a single composite criterion meant to represent "overall" job performance, think carefully about how to weight the components that constitute the composite. Several methods are available that allow the composite to reflect stakeholders' policy valuations.
- Keep in mind the distinction between nominal weights and effective weights. Applying equal weights to a set of components will almost certainly result in unequal contributions of the components to the composite. In most instances involving rationally determined weights, alternate empirical weights need to be calculated and applied to the components to obtain effective weights that will produce the desired weighting indicated by the nominal weights.

Chapter 14. References

Campbell, J. P. (1986). *Improving the selection, classification, and utilization of Army enlisted personnel Annual report, 1986 fiscal year* (Report 813101). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Campbell, J. P. (Ed.) (1990). Project A: The U.S. Army selection and classification project (Special Issue). *Personnel Psychology*, *43*, 231-378.

Campbell, J. P., & Knapp, D. J. (2001). *Exploring the limits in personnel selection and classification*. Mahwah, NJ: Erlbaum.

Green, P. E., & Srinivasan, V. (1978). Conjoint analysis in consumer research: Issues and outlook. *Journal of Consumer Research*, *5*, 193-123.

Johnson, R. M. (1974). Trade-off analysis of consumer values. *Journal of Marketing Research*, *11*, 121-127.

Piña, M., Jr., Emerson, M. S., Leighton, D., & Cummings, W. (1988). *Processing and Classification of Enlistees (PACE) system payoff algorithm development* (AFHRL-TP-87-41). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Rogelberg, S. G., Ployhart, R. E., Balzer, W. K., & Yonker, R. D. (1999). Using policy capturing to examine tipping decisions. *Journal of Applied Social Psychology*, *29,* 2567-2590.

Sadacca, R., Campbell, J. P., White, L. A., & DiFazio, A. S. (1989). *Weighting criterion components to develop composite measures of job performance* (Report 838). Alexandria, VA: U.S. Army Research Institute.

Torgerson, W. S. (1958). *Theory and methods of scaling*. NY: Wiley.

Wang, M., & Stanley, J. (1970). Differential weighting: A review of methods and empirical studies. *Review of Educational Research, 40,* 663-705.

Ward, J. H., Jr. (1977, August). *Creating mathematical models of judgment processes: From policy-capturing to policy-specifying* (AFHRL-TR-77-47, AD-AO48 983). Brooks AFB, TX: Occupation and Manpower Research Division, Air Force Human Resources Laboratory.

Chapter 15.
Multiple Hurdles and the Correction for Range Restriction
Norman M. Abrahams

Introduction

The main objective of conducting a predictive validity study in the personnel selection context is to determine the value or effectiveness of the selection test in screening job applicants. However, many validation strategies do not take into account that there may be more than one selection stage in an organization's hiring process. That is, typically, we deal with a single decision point. Multiple hurdle selection systems have more than one. In this chapter, we describe several sequential testing or multiple hurdle selection systems and the technical problems that may lead to inaccurate estimates of test validity (downward bias) when a hurdle is not taken into account, or is inappropriately dealt with. We also explore some of the literature on multiple hurdles and several remedies, some of which involve the correction for range restriction formulas described in earlier chapters.

Multiple Hurdle Selection Systems

In general, the range restriction problem is not unique to the field of personnel selection, nor is the problems of accurately estimating the unknown unrestricted correlation when there have been violations of the underlying Pearson-Lawley assumptions, such as curvilinearity. For example, the range restriction problem arises in other fields such as health (e.g., participants in a clinical trial drop after the first phase because they did not reach a critical level of improvement), econometrics (e.g., there are unobservable lower wage earners in a wage comparison study involving all occupations), and education, where, for example, college applicants may be rejected because they do not meet an official score on the Scholastic Aptitude Test (SAT) or the Academic College Testing (ACT) (a range restriction situation similar to military qualification on the ASVAB). As another example, students who qualify for a high ranking university may not even bother to apply to a lower tier college resulting in restriction in range in high scores (see Sackett and Yang, 2000, for more types of restriction in range).

Besides these obvious cases of range restriction, there are others that are not so obvious. For example, a structure interview may be used in college admissions but the instrument may not be scored and so, not used in validity analysis. If the institution wants to estimate the validity of the SAT/ACT in predicting first year grade point average for the applicant population, the omission of the interview instrument scores may cause the range corrected validity of SAT/ACT to be downward biased. The use of a structured interview may also be used to eliminate candidates from employment consideration. Generally the interview is administered after a general cognitive test has screened applicant out and so acts as a second stage assessment, part of a sequential testing or multiple hurdle selection system.

Multiple hurdles, as opposed to multiple cutscores, are sequential screening/ selection systems that are formally structured in stages so that not all applicants are administered all of the tests or instruments. Essentially, an applicant must meet the cutscore on the first screen to progress to the second screen. For example, the Federal Aviation Administration (FAA) administers a general cognitive ability test to Air Traffic Controller applicants in an initial screening stage and those failing the test are eliminated from further employment consideration. In a second stage of testing, those passing this first stage of testing are administered a more expensive battery of comprehensive and time-consuming tests. Those applicants who survive both stages (and any other hurdles) are eligible for employment consideration. Multiple hurdle selection systems provide practical benefits to both the organization and the applicants. From the organizational perspective, the most expensive personnel selection procedures are conserved. From the applicants' perspective, time spent in a fruitless effort to gain employment is minimized, freeing those rejected in the first stage to pursue other employment opportunities.

There are at least two examples of multiple hurdle cognitive screening systems in the military involving the Armed Services Vocational Aptitude Battery (ASVAB). The first example is the use of a subset of the ASVAB tests, the ASVAB math and verbal tests—the Armed Forces Qualification Test AFQT) for military service eligibility (see Chapter 2 in the Introductory Manual). Of course, other military screens are applied that include education and non-cognitive screens such as physical fitness, medical, and moral status. If all of the military eligibility screens are passed (hurdles), the ASVAB is again used in a second stage of cognitive screening for occupation qualification. Technically, because all military applicants must take the full ASVAB for both enlistment and occupational classification, the ASVAB use can be thought of as a multiple cutoff system. The only differentiating factor that makes this screening system a hurdle and not a multiple cutscore system is that the selection and classification decisions are separate processes.

Another example of an ASVAB multiple hurdle situation that is often not recognized as such is when the criterion that the ASVAB is targeted to predict is job performance. The first hurdle, of course, is the ASVAB standard for military eligibility, but technically is not considered so because the scores for both enlistment and job classification are available for all applicants. So, we say the second hurdle screen is passing the training course to "qualify" for reporting to the job. Just as many military enlisted members do not meet the ASVAB standard for a specific occupation, many who do qualify and attend the training course do not meet the training standard and therefore fail the training and do not report to the job. In this multiple hurdle case, measures of training performance serve the same purpose as the ASVAB classification standard – to screen out individuals from the job who are at an unacceptable risk for failure. Ignoring a screening hurdle has analytical consequences when validating the ASVAB, discussed in the following sections.

Technical Issues with Multiple Hurdles

Earlier chapters introduced the Pearson-Lawley correction for range restriction formulas. We recall that the assumptions for applying the formulas for the simple bivariate case are (a) linearity in regression of y on x throughout the unrestricted

bivariate distribution and (b) homoscedasticity of y error variance conditional on the values of x. An additional assumption is that explicit selection has occurred only on x. We can consider the assumptions (a) and (b) as "distribution assumptions" and the explicit selection on x assumption as a "selection assumption." Lawley (1943) relaxed the distributional properties of x and y for the multivariate case (i.e., the formal properties of test score normality) but maintained that linearity and homoscedasticity should hold. For the multivariate case, the selection assumption is that y_i covariances are unconditional on x_i. As a reminder, the three-variable case of one explicit selection variable and two incidental variables (commonly, one explicit selector, one experimental predictor, and one criterion variable) is a specific case of the general multivariate formulas (see Chapter 5).

It is well recognized and thoroughly discussed in previous chapters that the underlying assumptions for correcting for range restriction in the two-variable (x,y) bivariate normal case is linearity of y regressed on x across the total x-score range with homoscedasticity of error variances. Bivariate normal distributions are not always attained, however, and it is left to the researcher to assess the state of the only partially observable data (due to a selection standard in our illustration). Much research has been conducted on violations of the linearity and homoscedasticity assumptions and why these violations occur (some discussed in Chapter 11). We refer the reader to Dunbar and Linn (1991), Linn (1983), and Sackett and Yang (2000) for graphical representations of the non-linearity relations between two variables that can occur when a third "selection" variable selection has not been accounted for. As *our* visual aid, we first refer to Figure 15-1 for the depiction of the relation between two hurdle variables, H_1 and H_2, before and after a cutscore has been applied to H_1.

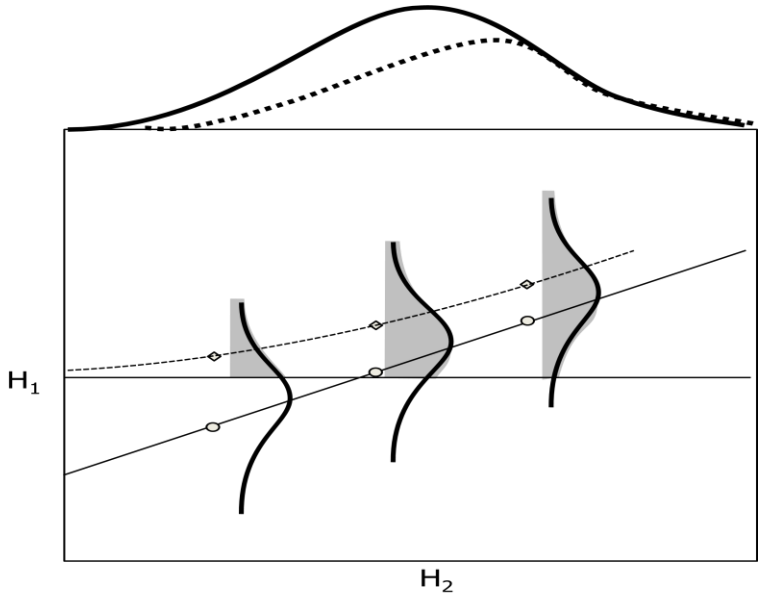


Figure 15-1. Non-linearity relation between two hurdle tests (H_1 and H_2) resulting from the H_1 cutscore.

Figure 15-1 shows the linear regression line (H_1 on H_2) before a cutscore has been applied to H_1, and also the slightly curvilinear line above that results from a cutscore applied to H_1. We can see clearly that the linear regression line (roughly drawn to go through the mean of the H_1 distribution observed at each H_2 score) becomes curvilinear because in the very low H_2 score range, most of the low scores on H_1 are eliminated due to the H_1, H_2 correlation. Obviously, to the extent that scores on H_2 are highly correlated to scores on the performance measure, say Y (now substituting for H_2), the curvilinear relationship between H_1 and Y also will be observed and the validity of H_1 in predicting Y will be downward biased (as the slope of the line is flattened). We could say for Figure 15-1 that H_1 is the ASVAB, H_2 is training grade, and Y is a measure of job performance. We can see how the ASVAB's predictive validity will be underestimated with job performance as the criterion if we ignore the training hurdle.

A Constructed Two-Hurdle Example

This section describes the technical aspects of estimating the validity of ignored, not recognized, or inappropriately dealt with hurdles. A hypothetical case was constructed in which we predetermined the correlations, means, and standard deviations for two hurdles (H_1 and H_2) and a performance criterion (Y) in an applicant population drawn from a trivariate normal distribution. For convenience, we limited this population to 1,000 cases. The three variables were standardized to have means of 0 and standard deviations of 1. In the first scenario, an inappropriately dealt with hurdle, H_1 is a recognized formal hurdle whereas H_2 is also a hurdle but ignored. The objective of our investigation was to compare the accuracy of the validity estimations for both H_1 and H_2 for the scenario where the true validity coefficients are known for both variables.

Table 15-1 shows the correlation matrix constructed for a hypothetical applicant population with means and standard deviations for the three variables just described.

Table 15-1
Hypothetical Applicant Population (N = 1,000) H_1, H_2, and Y
Means, Standard Deviations (SD), and Intercorrelations

	H_1	H_2	Y	Mean	SD
H_1	1.00	.45	.46	0.00	1.00
H_2		1.00	.53	0.00	1.00
Y			1.00	0.00	1.00

Next, we assume that the top 28% of applicants scoring highest on H_1 in Table 15-1 were administered H_2. The Y variable would not be known at this point. Table 15-2 shows the resulting H_1 and H_2 correlation matrix, determined analytically.

Table 15-2
Hurdle 1 Selectees (n = 280) with H_1 and H_2 Means, Standard Deviations, and Intercorrelations

	H_1	H_2	Mean	SD
H_1	1.00	.25	1.20	0.51
H_2		1.00	0.54	0.92

Table 15-2 shows that the range restriction that occurred by selection on H_1 attenuates the H_1H_2 correlation (initially .45 to a now much lower .25). Consistent with range restriction effects, the mean scores are higher for both H_1 and H_2 (than the initial standard score mean of 0) and the standard deviations lower (than the initial 1.0).

Finally, we assumed that 160 of the 280 applicants who scored highest on H_1 scored highest on H_2 (16% of the original applicant population). The 160 final selectees went on to the training program to be scored on the performance variable Y. Table 15-3 shows the resulting H_1, H_2, and Y correlation matrix from the fabricated two-hurdle selection system, also determined analytically.

Table 15-3
Hurdle 2 Selectees (n = 160) with H_1, H_2, and Y Means, Standard Deviations, and Intercorrelations

	H_1	H_2	Y	Mean	SD
H_1	1.00	.19	.23	1.28	0.54
H_2		1.00	.31	1.17	0.59
Y			1.00	0.83	0.87

Table 15-3 shows that the range restriction that occurred by selection on both H_1 H_2 further attenuates the H_1H_2 correlation (.25 in Table 15-2, after the first hurdle; .19 in Table 15-3, after the second hurdle). Consistent with increased range restriction effects, the mean scores are even higher for both H_1 and H_2, and at least for H_2, the standard deviation is lower. Further, because the Y variable, observed for the first time, is correlated with both H_1 and H_2 in the population (.46 and .53, respectively), these correlations are also reduced in magnitude (.23 and .31, respectively).

We remind ourselves that H_2 was erroneously thought to be an experimental predictor, not a formal selection instrument with an applied cutscore. As often occurs in the evaluation of an experimental predictor, we might assume that it is appropriate to apply a conventional Pearson-Lawley correction that treats H_2 as an incidental selection variable (Pearson's Case III, which is Equation 5-19 in Chapter 5 and Equation 7, p. 174, in Thorndike, 1949). This correction procedure is not technically appropriate, however, because H_2 *was* a formal explicit selection instrument. Table 15-4 shows the estimated applicant population matrix resulting from misapplication of the Case III formula.

Table 15-4
"Corrected" Applicant Population H_1, H_2, and Y Means, Standard Deviations, and Correlations

	H_1	H_2	Y	Mean	SD
H_1	1.00	.33	.39	0.00	1.00
H_2		1.00	.37	0.91	0.61
Y			1.00	0.37	0.92

The Table 15-4 values differ markedly from the known applicant population values in Table 15-1. The inappropriately used Case III formula produced corrections that not only underestimated the H_1 and H_2 applicant population validities, but also "reversed" their relative standings in predictive effectiveness. In the applicant population, H_2 has higher validity than H_1 (.53 vs. .46, respectively), but Pearson's Case III correction erroneously produced .37 versus .39, respectively.

Sequential use of the Pearson-Lawley Formulas

In two studies of multiple hurdles (only a part of which is reported in this chapter), Abrahams and Alf (1998; Alf & Abrahams, 1998) investigated and compared several variations of the Pearson-Lawley formulas for their potential to solve the multiple hurdle validity estimation problem. The authors found that applying the standard Pearson-Lawley formulas in a sequential manner yielded the exact population values (Table 15-1). Specifically, the H_1, H_2, Y matrix for the final selectee sample (Table 15-3) is used with the Pearson-Lawley multivariate formulas to estimate the missing Y for the Hurdle 2 matrix (Table 15-2) that contains only H_1 and H_2. In turn, the now complete three-variable Hurdle 2 matrix is used with the correction formulas to estimate the missing H_2 and Y applicant population data (Table 15-1).

Abrahams and Alf (1998; Alf & Abrahams, 1998) noted in their investigation of Lawley's (1943) original exposition of the multivariate correction procedures that a single statement supported the use of the sequential corrections. Specifically, Lawley noted that "Since the conditions (linearity and homoscedasticity) refer to relations existing between the y and x variables and take no account of the form of distribution of the x alone, it is clear that the selection formulas, if once applicable, may again be applied when a second selection is performed on the already selected population" (p. 29).

The use of the sequential Lawley correction assumes that data for those selected and rejected are available at each hurdle stage. Having complete data might not normally be the case. For example, those at the Hurdle 2 stage will have H_1 and H_2 scores, but those rejected on H_2 will not appear at the next stage where the Y measurement is taken. Another kind of "missing data" situation applies to those selected at both hurdles but some of whom fail on the criterion variable, in which case the failures' Y scores will be missing or, if entered, highly suspect. These two types of missing data must be imputed somehow to legitimately use the sequential Lawley correction method.

Abrahams, Alf, and Neumann (1993) provided an imputation procedure to compute *Y* performance scores that do not rely on predictor scores (i.e., *Y* scores for missing cases are not based upon regression equations that use the predictors). This imputation procedure has had application for scoring failures for large-scale ASVAB validation/ standards studies (Wise et al., 1992). However the procedure could just as well be applied to missing hurdle scores for those rejected at a hurdle and therefore not arriving at the stage where performance *Y* scores are taken. The "Scoring of Failures" procedure is based on the observed criterion distribution that is then anchored to the theoretically normal distribution. Because the method is not based upon regression analysis involving the predictors, the imputations may dampen the predictor validity coefficients. On the other hand, the procedure does have the potential to inflate validity estimates due to the direct tie to predictor scores (regression-based).

Some Psychometric and Econometric Methods

Alf and Abrahams (1998) explored not only the Pearson-Lawley formulas but also a variety of other correction methods (including Maximum Likelihood) for potential use in multiple hurdles selection systems. The methods can be categorized as either psychometric or econometric. The psychometric methods are variants of procedures developed by Pearson (1903) and Lawley (1943) and assume linearity of regression and homoscedasticity, as discussed earlier in this chapter and in previous chapters. The econometric methods (e.g., Heckman, 1979; Muthén & Joreskog, 1983) typically model the single-stage selection process developing a regression equation for y on x that includes a binary term to account for the range-restricted group. Typically, the numeric value of 1 is assigned to members of the unselected population, and 0 to the selected group. Some approaches use probit analysis as the first stage in the process of parameter estimation to account for the effects of sample selection.

The next step in a probit analysis is to include the probit values along with x in a least squares regression to predict y for all cases. The validity coefficient then becomes a straightforward calculation. The applicability of some of the econometric methods has, to a certain extent, been studied in psychometric settings. Nelson's (1984) research, however, suggests that the econometric methods are least effective where they are most needed—that is, where the sample selection is most stringent. In the case of selection stringency, the Pearson-Lawley procedures were more accurate than the econometric procedures. After reviewing several studies, Dunbar and Linn (1991) were not optimistic about the application of econometric methods to test validation methods. The relative ineffectiveness of the econometric methods may be due, in part, to their failure to use all of the available information available for use by the Pearson-Lawley procedures.

An important difference in the econometric and psychometric methods is that, although their intentions are identical, their assumptions differ. The econometric methods assume a strict cutscore or "threshold value" (as referred to by econometricians). *All* applicants below the cutscore are rejected and *all* above are selected. The psychometric methods, on the other hand, do not require the strict cutscore assumption and therefore are flexible in permitting the real-world possibility of a variety of reasons leading to selection or rejection. Further, the selection and rejection

can occur from any point in the predictor distribution. The restrictive assumption of the econometric methods, coupled with increased standard errors of their parameter estimates may, in combination, contribute to their relative ineffectiveness compared to the Pearson-Lawley methods.

It is important to note that *neither* the econometric nor psychometric approaches (without sequential corrections) deal specifically with evaluating multiple hurdle selection systems. For example, Linn (1983) demonstrated the inadequacies of both models in estimating conditional means and variances as a result of an unaccounted hurdle that resulted in a curvilinear regression of y on x.

Concluding Remarks

The importance of applying the appropriate procedures for correcting for range restriction in multiple hurdle selection systems cannot be overstated. There are real consequences in the applied setting for either not acknowledging a hurdle or inappropriately estimating a hurdle instrument's validity. The consequences of misestimating the validity of multiple hurdle selection instruments will most likely be reflected in an inappropriate setting of cutscores with a corresponding negative impact on the organization. The next chapter provides information about mainstream methods for dealing with multiple hurdle data considered within the framework of a missing data problem.

Chapter 15. References

Abrahams, N. M., & Alf, E. F. (1998). *A tool to optimize the predictive accuracy of personnel selection and classification instruments: Estimation of test validity in multiple hurdles selection.* (Contract Number: N66001-97-7903 CDRL A002AB). Final Phase II report prepared for Navy Personnel Research and Development Center, San Diego, CA.

Abrahams, N. M., Alf, E. F. Jr., & Neumann, I. (1993). The treatment of failures in validation research. *Military Psychology, 5,* 235-249.

Alf, E. F., & Abrahams, N. M. (1998). *A tool to optimize the predictive accuracy of personnel selection and classification instruments: Estimation of test validity in multiple hurdles selection.* (Contract Number: N66001-97-7903 CDRL A002AA). Final Phase I report prepared for Navy Personnel Research and Development Center, San Diego, CA.

Dunbar, S. B., & Linn, R. L. (1991). Range restriction adjustments. In A. K. Wigdor & B. F. Green (Eds.), *Performance assessment for the workplace*, *Vol II - Technical issues* (pp. 127-157). Washington, DC: National Academy Press.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica, 47,* 153-161.

Lawley, D. (1943). A note on Karl Pearson's selection formula. *Royal Society of Edinburgh, Proceedings, Section A, 62,* 28-30.

Linn, R. L. (1983). Pearson's selection formulas: Implications for studies of predicted bias and estimates of educational effects in selected samples. *Journal of Educational Measurement, 20,* 1-15.

Muthén, B., & Joreskog, K. G. (1983). Selectivity problems in quasi-experimental studies. *Evaluation Review, 7,* 139-174.

Nelson, F. D. (1984). Efficiency of the two-step estimator for models with endogenous sample selection. *Journal of Econometrics, 24,* 181-196.

Pearson, K. (1903). Mathematical contributions to the theory of evolution - XI. On the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions of the Royal Society, London, Series A,* 200, 1-66.

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology, 85,* 112-118.

Thorndike, R. L. (1949). *Personnel selection: Tests and measurement techniques*. NY: Wiley.

Wise, L., Welsh, J., Grafton, F., Foley, P., Earles, J., Sawin, L., & Divgi, D. R. (1992). *Sensitivity and fairness of the Armed Services Vocational Aptitude Battery (ASVAB) technical composite (*DMDC Technical Report 92-002). Seaside, CA: Defense Manpower Data Center.

Chapter 16.
Multiple Hurdles as a Missing Data Problem
Jorge L. Mendoza and Janet D. Held

Introduction

In the previous chapter, we were able to visualize and understand the downward bias in the validity coefficient resulting from ignoring a selection hurdle. We also saw that when all of the required data are available, the Pearson-Lawley procedures can be used in a sequential process starting from second stage data working back to the first stage. In this chapter, we take the position that a multiple hurdle selection system can be placed within the general framework of the missing data problem. Procedures such as Maximum Likelihood (ML) and Multiple Imputation (MI) are discussed, which have additional benefits over Pearson-Lawley in that the methods provide standard errors of prediction. For a full discussion of the problems in the context of selection test validation, we refer the reader to Dunbar and Linn (1991). The broad topic of missing data theory is more fully discussed by others (e.g., Little & Rubin, 2002; Rubin, 1976, 1996; Schafer, 2000, Schafer & Graham, 2002).

Some Missing Data Terminology

Missing Completely at Random (MCAR)

Data that are missing completely at random (MCAR) are exactly that—data that are not observed due to chance alone. MCAR data result from completely random processes (e.g., the inability or neglect of a data recorder). We do not address MCAR data in the single or multiple hurdles selection design other than to say they should not influence the regression of Y or X or the correlation between the two variables because there is no systematic pattern of missingness.

Missing at Random (MAR) and Missing Not at Random (MNAR)

Data that are missing at random (MAR) arise from the situation when performance Y scores are missing *strictly* due to selection on X, *assuming X and Y are correlated to some degree*. In personnel selection, factors such as applicant self-selection or rejection of job offers are factors that do not correlate with the reasons why some of the Y scores are missing. If they were correlated, then the situation would be missing not at random, or MNAR. For example, a MNAR situation would be when a selection process was based solely on a cognitive measure but job performance was related not only to cognition, but also to being able to get along with others. Further assume that those who cannot "get along" with others were fired. This situation would yield MNAR data, because some of the job performance Y scores are missing due to not being able to get along – that is, a factor that is related to Y scores and cannot be eliminated by controlling for the cognitively based X.

In personnel selection and the conduct of validity analysis, we are hoping for a MAR situation and we note that a MAR assumption is consistent with the Pearson-Lawley assumption of selection having occurred solely on *X*. Missing not at random (MNAR) means that some other selection mechanism besides *X* is responsible for the missing *Y* scores. This "other selection mechanism" might or might not correlate with *X* but must correlate with *Y* after controlling for *X*, thereby biasing the validity correction. MAR is a confusing term, because we intuitively think of the term "random" to mean just that – missing completely at random (MCAR).

Monotone Data

A monotone pattern of missing data means that all individuals having *Y* scores also have *X* scores. In a multiple hurdle selection system, we have labeled *X* as H_1, which is followed by H_2. In the multiple hurdle case, all individuals with *Y* scores will also have H_2 scores, and all individuals with H_2 scores will also have H_1 scores. A monotone data pattern is typically depicted as a set of stair steps, each decreasing in height going from left to right (e.g., see Mendoza, Munford, Bart, & Siew, 2004 and their stair step depiction for various test validation designs with ML estimation using the Estimation-Maximization Algorithm). Monotone patterns of missingness are a special case of the MAR assumption.

Ignorable and Non-ignorable Missingness

The ignorable missingness assumption (Rubin, 1976) means that the selection mechanism is known and the data are available. Ignorable missingness, consistent with MAR, is also equivalent to the Pearson-Lawley assumption of selection having occurred solely on *X*. The selection situation is ignorable in that the *Y* data are missing strictly due to the correlation of *X* with *Y*, not due to some unobserved selection mechanism that is correlated with *Y* after controlling for *X*. The ignorable selection situation applies to the multiple hurdle selection system when all hurdles are accounted for; if not, there is non-ignorable missingness. Adjusting regressions and correlations for non-ignorable missingness (such as the unaccounted-for institutional decisions based upon factors that correlated with both *X* and *Y*, and self-selection decisions that eliminate high aptitude/achievement youth from some low-tier colleges) affects the unrestricted population regressions and correlations. Both ignorable and non-ignorable missingness can be consistent with a monotonic data pattern; however, non-ignorable missingness is consistent not with MAR, but with MNAR. The terms MAR and "ignorable" are used interchangeably in the literature.

Censored Data

The term *censoring* is used in several fields to describe data that are missing due to the upper or lower limits of the measurement instrument, such as what psychologists would observe when an aptitude test has a ceiling effect, or what an economist would observe when an aptitude standard causes range restriction. The concept of censored data is not the same as truncated data, which introduces some confusion in our

personnel selection test validation designs. For example, ASVAB composite scores will be left truncated in a Navy school's data set due to a cutscore applied to the Navy Rating's operational ASVAB classification composite, whereas other ASVAB composite scores will be censored because all ASVAB composites are correlated (to varying extents). In either case, direct (explicit) selection or incidental (indirect) selection, we can deal with the downward bias in ASVAB validity coefficients when estimating them for the unrestricted population with the traditional formulas presented in Chapter 5.

Heckman (1976, 1979), as an economist, refers to censored data in the broadest of terms in that the data can be censored due to either an organization's decision-making process (some of which might or might not be accounted for) or a self-selection process (that cannot be accounted for). Censoring, unless purely random, violates the MAR assumption and the data are therefore MNAR. Heckman represents a MNAR selection model by depicting an organization's decision-making process that is not taken into account (our unaccounted for hurdle depicted in Chapter 14).

Addressing the Non-Ignorable Missing Data Problem

Researchers in academia have long been concerned with the non-ignorable missingness problem. For example, Linn (1968) pointed out the problem of establishing the validity of a career guidance test battery in making future decisions (e.g., further education or career choices) when those who take the battery are self-selected. Linn provided valuable insights into the problem:

> "In situations such as those encountered in attempting to validate a guidance test battery, the nature of the process of self-selection is very difficult to model and the true explicit selection variables are difficult to identify and/or to measure. It may be that the most reasonable approach to the problem of correcting for bias due to selection in such cases is to include as many measures which are thought to have relevance for the selection processes as is reasonable within the practical constraints of the situation" (p. 72).

More recently, Ryan, Sacco, McFarland, and Kriska (2000) pointed to such self-selection factors as an applicant's perceptions of the organization and motivation to obtain the job, as well as employment alternatives or offers from higher tier colleges. All of these self-selection factors are considered non-ignorable selection mechanisms that are difficult to model. The following subsections provide brief discussions about the salient points of research attempting to solve the standard range restriction problem and the non-ignorable missingness problem, some of which make comparisons to the Pearson-Lawley method.

Muthén and Yang Hsu (1993)

Muthén and Yang Hsu (1993) were concerned about the missing data problem in college or graduate school admissions. Muthén and Yang Hsu were most interested, however, in using structural equation modeling to find the structure and correlations

between the underlying latent factors in a test battery (not relations from observable test scores). Muthén and Yang Hsu pointed out that the Pearson-Lawley approach is not designed to deal with latent variables. (In simple situations, we can achieve similar results by correcting for the Pearson-Lawley estimators for unreliability.) We do not discuss the complexities of the study or findings here other than to say that the Pearson-Lawley and ML procedure are comparable under multivariate normality and full data availability. The benefits of the ML procedures are that they are theoretically supported and in principle are more flexible than the Pearson-Lawley corrections. Another advantage of the ML approach is that in many situations it gives estimates of standard errors that apply to the full information maximum likelihood using the observed information matrix. We refer the reader to Muthén and Yang Hsu or Enders (2010) for a full discussion of the ML procedures.

Mendoza, Bard, Mumford, and Siew (2004)

Mendoza et al. (2004) provided an algebraic extension of Pearson-Lawley. They also pointed out that the multiple-hurdle situation in many contexts is a monotone missing data problem and that the algebraic extensions under the assumption of multivariate normality and ignorable data provide ML estimates. The ML estimates can be more easily obtained using the EM (estimation maximization) algorithm. Although the EM algorithm provides ML estimates, it does not provide standard errors because it does not find the derivative in the maximization of the ML function. When the EM algorithm is used to find the ML estimates, the standard errors can be obtained using one of the many available bootstrap methods. Mendoza et al. covered three test validation designs (concurrent, predictive, and multiple-hurdle) and their EM solutions under the assumption of MAR and an ignorable selection mechanism. (The authors refer us to Sackett & Yang, 2000, for a complete taxonomy of correction procedures.)

For context, Mendoza et al. (2004) provided a citation for a multiple-hurdle selection system used to employ airport screeners (Kolmstetter, 2003). In the first stage, an online application was administered. Those who passed the first stage completed a computer-administered test battery in the second stage. A third stage consisted of a structured interview, followed by a physical ability test and a medical evaluation. The final stage was a security background check.

Three methods were applied to estimate (recover) the known unrestricted regression parameters and variable correlations: (a) the appropriate sequential correction formulas developed and described in the Mendoza et al. (2004) study, (b) the ML Expectation Maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977), and (c) Bayesian multiple imputation (MI) (Rubin, 1976). The results showed that the MI method was most accurate in estimating the known unrestricted regressions and correlations in the limited context when compared to the EM algorithm and the formulas, but that all three procedures (given large samples from a multivariate normal population) produce similar results. We note that the formulas developed by Mendoza et al. are extensions of Rubin's missing data algorithm and yield comparable results in the multiple-hurdle context that applied to the Pearson-Lawley sequential corrections.

Mendoza et al. (2004) took the position that even under the ignorable selection situation, there are some advantages to using either the EM or MI approaches over the formulas. Formula use requires exact specification of the test validation design, which is a function of how many hurdles are present, whereas the EM or Full Information Maximum Likelihood (FIML) procedures do not. Also, the EM algorithm "…yields ML estimators under a variety of missing data structures and is not limited to the monotonic missing data structure…" (p. 430). The MI procedure has the added advantage over EM in that it provides a facsimile to a formula-based standard error. Mendoza suggested that procedures that augment EM with bootstrap standard errors may also be useful for conducting test validation. We refer the reader to some of the bootstrap literature that suggests accurate standard errors of the range-corrected correlation coefficient (Chan & Chan, 2004; Li, Chan, & Cui, 2010; Mendoza, Hart, & Powell, 1991).

Olson and Becker (1983)

Olson and Becker (1983) graphically portray another non-ignorable selection situation that applies to the military setting where many high aptitude/achievement youths opt for college rather than the military service (particularly in times when there is a large supply of high paying private sector jobs that require a college education). Figure 16-1 is the opposite situation displayed by Figure 15-1 (Chapter 15) where low aptitude/ability individuals are screened out of the selection process.

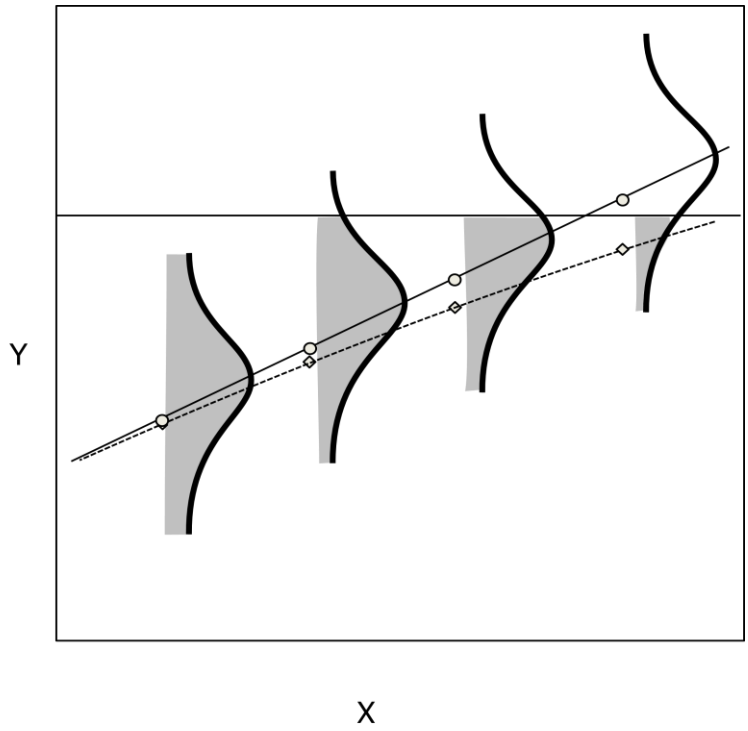


Figure 16-1. Non-linearity and heteroscedasticity resulting when high aptitude youth select out.

Figure 16-1 represents a regression situation (say the ASVAB as *X* and final school grade as *Y*). Assuming a substantial correlation between *X* and *Y*, there will be missing or sparse data in the upper *X,Y* score range if many high ASVAB youth chose college or high paying jobs over military service. The missingness effect will be a downward bias in the population regressions and correlation. We note for Figure 16-1 that the *X* cutscore is not shown and that the displayed *Y* cutscore is notional not reflecting accuracy or at what point the line would cross the *y*-axis.[14] Olson and Becker (1983) also noted that complete truncation on *Y* is not in reality a situation experienced by organizations, and psychologists' formula to correct for complete *Y* truncation, presented by Thorndike (1949) as his Case 1 (p. 173), would never be used in application. Further, the Thorndike Case 2 solution (p. 173) that corrects for incidental selection on *Y* due to explicit selection on *X* is not appropriate because in a nonignorable selection situation (either low or high score missingness), the explicit *X* variable in the equation is not the only explicit selection mechanism.

Organizations that conduct utility analysis to evaluate their personnel selection systems should realize that if a non-ignorable selection mechanism is not accounted for (either low or high score missingness), there may be under prediction of the performance Y scores due to smaller regression weights (a lower slope). We refer to the quotation by Linn (1968) about the need to include all variables that are relevant in the estimation of a selection test's validity coefficient (this chapter) and add a partial quote from Thorndike provided by Olson and Becker (1983).

> "When selection is based, as it often is, on a clinical judgment which combines in an unspecified and inconstant fashion various types of data about the applicant, and when this judgment is not expressed in any type of quantitative score, one is at a loss as to how to estimate the extent to which the validity coefficient for any test procedure has been affected by that screen" (Thorndike, 1949, p. 176).

Olson and Becker proposed, as others have, an analytical method based on the econometric literature to address the special case of "…omitted variable bias" (p. 143). We discuss the method in the next section as developed by Heckman (1976, 1979) (mentioned in Chapter 15) and applied to our more familiar academic selection context.

Gross and McGanney (1987)

Gross and McGanney (1987) were concerned about the non-ignorable missing data problem in the academic setting and the shortfalls of the traditional correction formulas (i.e., Pearson-Lawley). "The traditional correction formula approach can be viewed as the special case of the general model in which one assumes a priori no interrelation" (p. 605). Gross and McGanney acknowledged that the Heckman (1976, 1979) econometric missing data model could be used as part of a solution. The Heckman model, compared

[14] We refer the reader to Chapter 3 about the cautionary note from Smith (1948) on using the Taylor-Russell (1939) tables when the assumption of bivariate normality in the *X*/*Y* relation is not upheld.

to the Pearson-Lawley (traditional correction formulas), allows specification of the relation between the non-ignorable selection process and *Y*. Gross and McGanney described a potentially useful statistical model that has three components:

> "... (a) a regression model that expresses the *xy* relation, (b) a selection model that describes the selection process (i.e., the basis for the missing *y* scores), and (c) an assumption concerning the relation between the two former models" (p. 605).

Gross and McGanney (1987) started their discussion of the first model component with the basic population-based regression model,

$$y = B_0 + B_1 x + e_y \qquad (16\text{-}1)$$

and pointed out that the correlation regression parameters $p(x,y)$ can be expressed from regression analysis terms (their Equation 2) as:

$$p(x, y) = \left[\beta_1 \cdot \sigma(x)\right] / \left[(\beta_1 \cdot \sigma(x))^2 + \sigma^2(y \mid x)\right]^{1/2}. \qquad (16\text{-}2)$$

Gross and McGanney (1987) started their discussion of the *second* model component, that is, the process of selection that explains observable and unobservable *y* scores, by defining y_s as observable only if a threshold value is exceeded on the selection mechanism. A regression model for y_s that parallels Equation 16-1 (their Equation 3) is,

$$y_s = \alpha_0 + \underline{\alpha}' \cdot \underline{x}_s + e_s \qquad (16\text{-}3)$$

where $\underline{x}_s$, the actual selection variable(s), could be (a) the same as the predictor *x* variable of interest, (b) other than the *x* variable, or (c) a mix of both. Given some distribution normality assumptions that Gross and McGanney, an equation can be developed (shown as their Equation 4, the normal ogive or probit function) that estimates the probability of selection and thus has an observed *y*, given $\underline{x}_s$. The variable y_s is therefore only partially observable to the extent that all of the $\underline{x}_s$ are observed.

Gross and McGanney (1987) started their discussion of the *third* component of the model, the relation between the regression and selection models, with an assumption of bivariate normality between *x* and $\underline{x}_s$, and between *y* and y_s, with the primary interest in the correlation between *y* and y_s conditional on *x* and $\underline{x}_s$, $p(y, y_s|x, x_s)$. The authors state "This correlation is a key parameter of the model because it determines whether the selection process is ignorable (i.e., whether there is a relation between *y* and the probability of selection)" (p. 606). That is, if $[p(y,y_s|x, x_s) = 0]$, then the traditional Pearson-Lawley correction will suffice (to the extent that all underlying assumptions for performing the correction apply). Conversely, if the yy_s correlation is not zero (y_s modeled as a latent variable), then there has been a latent unobserved selection mechanism and an additional regression parameter must be included in the corrected validity estimation model.

Gross and McGanney (1987) evaluated two procedures for estimating the unrestricted *x,y* correlation, a two-step process (Heckman, 1976, 1979; Olsen & Becker, 1983) and ML (comparing the results with the Pearson-Lawley range correction). We refer the interested reader to the article for the details and merely mention that there appeared to be mixed results but favoring the ML procedure. However, the authors in their summary section expressed general concerns about the adequacy of sample size, potential multicollinearity issues between the x and x_s variables, the bivariate normality assumption for y and y_s, and the possibility that high ability individuals as well as low ability may be missing from the data due to nonignorable selection processes.

Concluding Remarks

We refer the reader to Enders (2010) for an applied approach to dealing with missing data and full discussions of the newer methods considered as state-of-the-art. Enders provided an employee selection data set that he used to demonstrate the pros and cons of the various methods. He also discussed software packages and provided a list of recommended readings after each chapter and a website where some syntax can be obtained (www.appliedmissingdata.com). However, Enders reminded us of several important points: "A missing data handling technique is only as good as the veracity of its assumptions…" and "Until more robust MNAR analysis models become available (and that may never happen), increasing the sophistication level of the MAR analysis may be the best we can do" (p. 344). We also refer the interested reader to the recommendations of The National Academies Panel on Handling Missing Data in Clinical Trials (National Research Council, 2010). First, the panel emphasized the role of design to limit the amount and impact of missing data. Two of their 18 recommendations were of special interest to personnel researchers.

- Recommendation 3: "Trial sponsors should continue to collect information on key outcomes on participants who discontinue their protocol…" (p. 3), and
- Recommendation 15: "Sensitivity analyses should be part of the primary reporting of findings from clinical trials. Examining the sensitivity of the assumptions about the missing data mechanism should be a mandatory component of reporting" (p. 5).

These two recommendations echo the comments made here and elsewhere about the importance of collecting additional data that could be helpful in understanding missingness, and also the importance of carrying out several different analyses if the MAR assumption is suspected.

Chapter 16. References

Chan, W., & Chan, D. W. –L. (2004). The bootstrap standard error and confidence intervals for the correlations corrected for range restriction: A simulation study. *Psychological Methods, 9*, 369-385.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39,* 1-38.

Dunbar, S. B., & Linn, R. L. (1991). Range restriction adjustments. In A. K. Wigdor & B. F. Green (Eds.), *Performance assessment for the workplace, Vol II - Technical issues* (pp. 127-157). Washington, DC: National Academy Press.

Enders, C. K. (2010). *Applied missing data analysis*. NY: *Guilford*.

Gross, A. L., & McGanney, M. L. (1987). The restriction in range problem and non-ignorable selection processes. *Journal of Applied Psychology, 72, 604-610*.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection, and limited dependent variables, and a simple estimator for such models. *Annals of Economic and Social Measurement, 5,* 475-492.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica, 47,* 153-161.

Kolmstetter, E. (2003). I-Os making an impact: TSA transportation security screener skill standards, selection system, and hiring process. *Industrial-Organizational Psychologist, 40, 39-46*.

Li, J. C., Chan, W., & Cui, Y. (2010). Bootstrap standard error and confidence intervals for the correlations corrected for indirect range restriction. *British Journal of Mathematical and Statistical Psychology, 64,* 1-21.

Linn, R. L. (1968). Range restriction problems in the use of self-selected groups for test validation. *Psychological Bulletin, 69,* 69-73.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (Rev. ed.). NY: Wiley.

Mendoza, J. L., Bard, D. E., Mumford, M. D., & Siew, A. C. (2004). Criterion-related validity in multiple-hurdle designs: Estimation and bias. *Organizational Research Methods, 7,* 418-444.

Mendoza, J. L., Hart, D. L., & Powell, A. (1991). A bootstrap confidence interval based on a correlation corrected for range restriction. *Multivariate Behavioral Research, 26,* 255-269.

Muthén, B. O., & Hsu, J. Y. (1993). Selection and predictive validity with latent variable structures. *British Journal of Mathematical and Statistical Psychology, 46,* 255-271.

National Research Council. (2010). *The prevention and treatment of missing data in clinical trials*. Panel on Handling Missing Data in Clinical Trials. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Olson, C. A., & Becker, B. E. (1983). A proposed technique for the treatment of restriction of range in selection validation. *Psychological Bulletin, 93,* 137-148.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63,* 581-592.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association, 91,* 473-489.

Ryan, A. M., Sacco, J. M., McFarland, L. A., & Kriska, D. S. (2000). Applicant self-selection: Correlates of withdrawal from a multiple hurdle process. *Journal of Applied Psychology, 85,* 163-179.

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology, 85,* 112-118.

Schafer, J. L. (2000). *Analysis of incomplete multivariate data*. Boca Raton, FL: Chapman & Hall/CRC.

Schafer, J. L., & Graham J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods 2002, 7,* 147–177.

Thorndike, R. L. (1949). *Personnel selection: Tests and measurement techniques*. NY: Wiley.

Chapter 17
Setting ASVAB Cutscores
Janet D. Held

Introduction

Chapter 3 introduced us to the various ways we can interpret the correlation (validity) coefficient. One of those ways is obviously tied to the magnitude of the validity coefficient that applies to a given selection instrument and, as a result, the extent to which we can improve a given inadequate success rate by raising that selection instrument's cutscore. Chapter 3 showed us how to construct an empirically-based expectancy (cutscore) table and also the theory-based Taylor-Russell (1939) tables. When conducting ASVAB validation/standards studies, we should recognize that there are limitations with empirical-based expectancy tables in that they are only appropriate for the operational selection/classification instrument upon which explicit selection has taken place and not for a candidate replacement because there is a floor of aptitude/ability already established due to the operational standard. Also, the empirical expectancy analysis cannot be used to assess performance impact from *lowering* the operational cutscore because individuals with scores below the cutscore would not be qualified. If such data points were observed below the cutscore, ASVAB waivers would have been given and we would not know the basis of the waiver decision (and even if we did, the sample size would be typically too small to include in an analysis). In this chapter on cutscore setting, we assume that (a) point waivers are not considered in cutscore analysis (although we provide guidance later in the chapter) and (b) the estimate of the population validity of a selection instrument is accurate.

Three Approaches in the Literature to Setting Cutscores

Two common approaches to setting cutscores for hiring decisions discussed in the literature are banding and top-down selection (e.g., Aguinis, 2004; Truxillo, Donahue, & Sulzer, 1996). The banding approach explicitly recognizes that there is measurement error in everyone's observed test scores. Banding undifferentiates individuals who score slightly lower or higher on a selection instrument's cutscore and therefore affords opportunity to consider, or place more emphasis on, other selection factors, such as community service, education, job experience, race/ethnicity, etc.

In contrast, the top-down approach takes the position that hiring candidates with the highest scores on the selection instrument will benefit the organization in terms of optimal job performance (see Sackett & Roth, 1991 for a comparison of the top-down and banding cutscore approaches). In the military enlistment context, a top-down ASVAB score approach is never taken and, in fact, the Navy's operational classification algorithm has a curtailment component for largely overqualified recruits for specific Ratings (discussed in the next chapter).

Besides the banding and top-down approaches to setting cutscores, a third approach is to use experts' judgments to come to a cutscore consensus: a content-based approach as opposed to a criterion-based approach. The Angoff (1971) method is most frequently evaluated against other content-oriented approaches (e.g., Truxillo, Donahue, & Sulzer, 1996) and involves deciding what constitutes a "minimally qualified" individual. In the military context, meeting training objectives constitutes a minimally-qualified graduate (See Chapter 5 in the Introductory Manual for information on Navy training).

Cascio and Aguinis (2005) stated that "It is unrealistic to expect that there is a single 'best' method of setting cutoff scores for all situations" (p. 227). We address other approaches and situations more in-depth in the following sections.

The Approach of Minimizing Classification Decision Errors

One also can set cutscores by deciding to minimize classification errors. Ghiselli, Campbell, and Zedeck (1981) described two classic examples (e.g., Cureton, 1957) for setting cutscores (p. 308) that would in every case minimize the two false classification errors: accepting those who would have failed and rejecting those would have passed. Both types of classification errors are of concern to the military but the emphasis shifts with changing recruiting environments (e.g., enlisting youth having a high risk of failing may be a tolerable position in a difficult recruiting environment if many would still be expected to succeed). Figure 17-1 shows two score distributions of one selection instrument - one for Failures and one for Successes - and the optimal cutscore that minimizes the two classification decision errors.

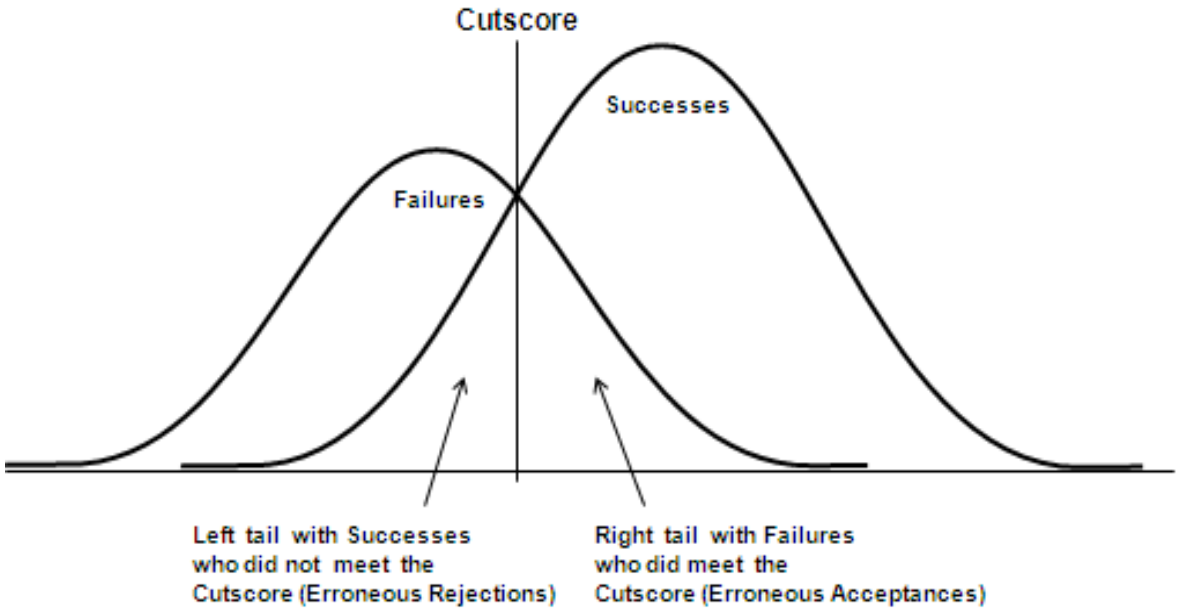


Figure 17-1. Optimal cutscore for minimizing classification decision errors.

The cutscore is set in Figure 17-1 so that it intersects the two distributions at the point at which the right tail for Failures and left tail for Successes intersect. The reason that any other cutscore would not minimize the sum of the two classification decision errors is that the tails of a normal distribution contain relatively fewer cases than at any other segment demarcated by an equal interval test score distance. This fact is readily seen by the normal test score distribution depicted in Figure 17-2.

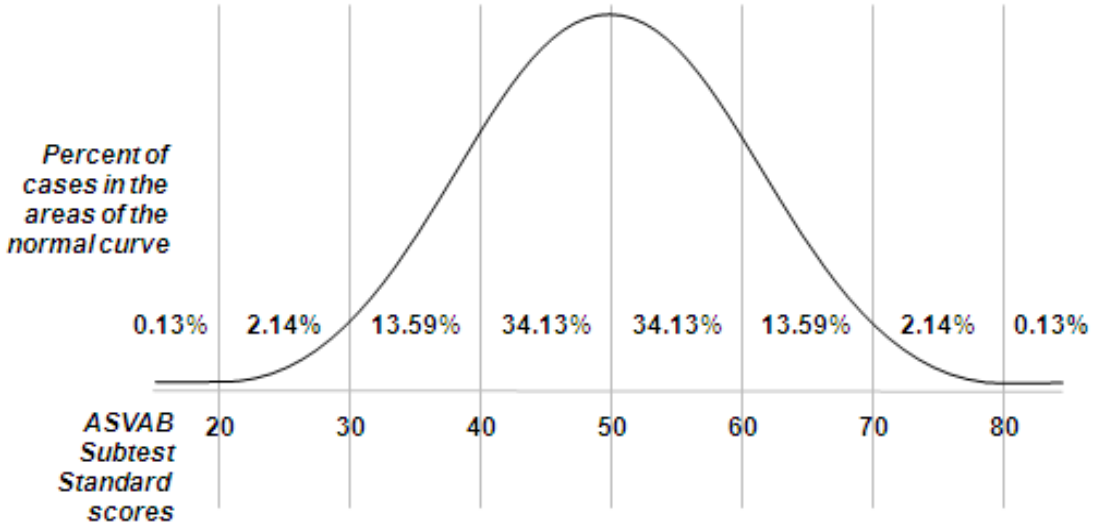


Figure 17-2. Partitioned areas under the normal curve.

With reference to Figure 17-1 and Ghiselli et al. (1981), any cutscore adjustment that moves either to the left or right would either be ascending or descending a "hill." In either case, the gain in cases from ascending the hill would be greater than the loss of cases to the counterpart tail. Therefore the cutscore intersecting the two distributions minimizes classification decision errors, no matter the extent of overlap in the Success and Failure distributions. As can be seen, this and the three previous approaches discussed about setting cutscores do not consider the correlation coefficient. Ghiselli et al. noted that in diverting attention from the correlation coefficient to, in essence, prediction error, the method "…makes no assumption about the form of the distribution and takes advantage of the nonlinear as well as linear components of association between the two variables" (p. 310).

There are two major issues with the "minimizing classification error" approach to setting ASVAB cutscores that make it inapplicable for the military. First, it is always the case that the military must trade off recruiting and training resources, and so the value of reducing each type of error would never be equal and may quickly reverse in changing recruiting environments. Second, there is more than one occupation to consider in a total cutscore system, which adds another dynamic - an occupation's value to the Navy (i.e., lowering the cutscore for one Navy Rating so that the cutscore can be raised for another).

Anchoring Cutscores to the ASVAB Normative Distribution

In Chapter 7 of the Introductory Manual (Synthetic Validity), we discussed the situation where a new Navy Rating is stood up and school performance data are not yet available to conduct validity or cutscore analyses. If the ASVAB validation/standards team is fully versed in the difficulty of training and graduation rates across Navy Ratings and also, the ASVAB validity, an educated/informed decision can be made about an interim ASVAB standard. One can gauge an ASVAB standard by anchoring it on a generic normal curve relative to relevant Ratings' ASVAB standards, as depicted in Figure 17-3.

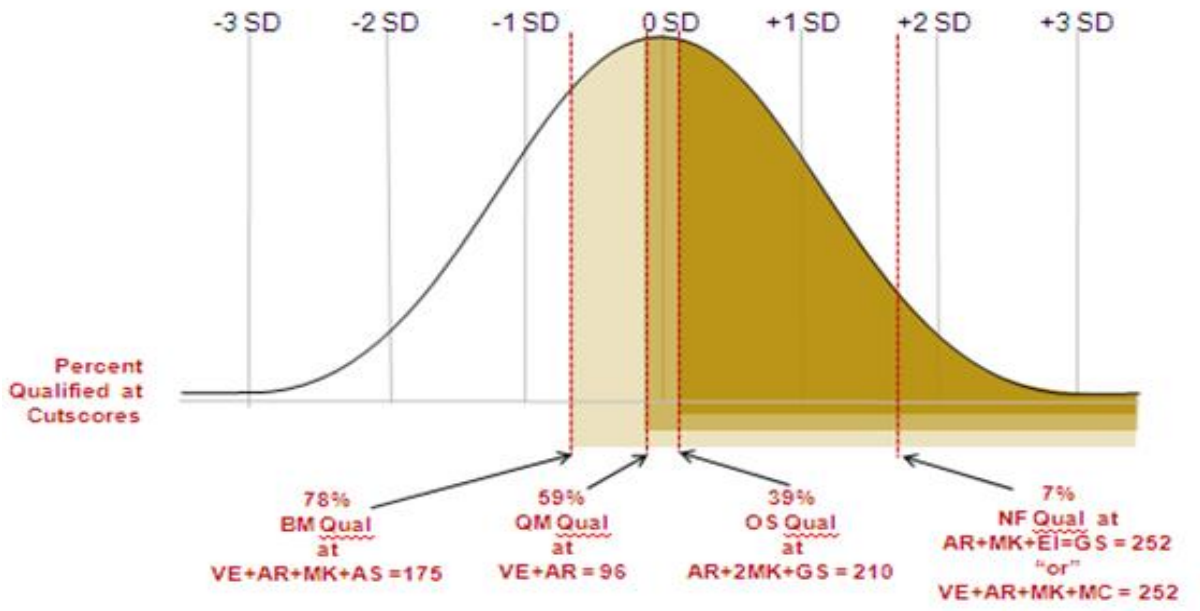


Figure 17-3. Four Ratings' ASVAB cutscores positioned on a normal test score distribution.

Figure 17-3 shows a rough placement of four Ratings' ASVAB standards (composites with cutscores) on a normal test score curve notionally applying to the PAY97 ASVAB normative population (Segall, 2004). The Ratings are Boatswain's Mate (BM), Quartermaster (QM), Operations Specialist (OS), and Nuclear Field (NF), which actually comprises three Ratings (see Appendix B of the Introductory Manual). The underlying principle for using the normal curve approach is that, even though the ASVAB composites differ in composition, the cutscore on each reflects the same *level* of aptitude/ability/skill/experience. Setting an ASVAB standard in this manner assumes that we understand (a) the underlying constructs measured by the composite and the linkage to the training content, (b) the expected or known difficulty of the curriculum and time allowed to train, all assessed relative to other Ratings that have operational ASVAB standards that can be evaluated for effectiveness, and (c) that eventually we will be able to confirm the effectiveness of the standard we operationalize.

Cutscores using the Taylor-Russell (1939) Tables

In Chapter 3 (about interpreting the validity coefficient), we discussed the Taylor-Russell (1939) tables and how they can be used in theoretically-based cutscore analysis. (Chapter 3 also provides particulars about the empirically-based cutscore table.) The way the Navy ASVAB validation/standards team uses the Taylor-Russell tables requires an estimate of the unrestricted in range validity coefficient that applies to the PAY97 ASVAB normative population (Segall, 2004) because the tables are based on bivariate normal distributions and the PAY97 scores are full range.

We also saw in Chapter 3 that the expected improvement in the success rate (training graduation rate in our case) from raising a cutscore is tied to, all other things being equal, the magnitude of the validity coefficient. In this section we illustrate how the tables can be used to assess the expected positive and *negative* impact on success rate from cutscore adjustments. Table 17-1 is taken from the Nuclear Field (NF) study (Appendix B in the Introductory Manual) and is used here to illustrate some important points about the use of the Taylor-Russell (1939) tables.

Table 17-1
Drop in Taylor-Russell (1939) Table Success Rates When Predictor Unreliability Lowers ASVAB Validity

Taylor Russell Base Rate Table	Validity	VE+AR+MK+MC Cutscore Applied to the ASVAB Normative Youth Population				
		257	245	236	229	223
		Qual(SR) = .05	Qual(SR) = .10	Qual(SR) = .15	Qual(SR) = .20	Qual(SR) = .25
.15	.85	.88	.76	.66	.58	.51
.20	**.80**	**.89**	**.79**	**.71**	**.65**	**.59**
.25	.75	.89	.81	.74	.69	.64
.30	.70	.89	.82	.76	.72	.67

Note. The base rate is the success rate that would be observed if all individuals in the population of interest were selected for a job (or training, in our case) without applying an aptitude standard. Each line in our table references a different Taylor Russell base rate table. Qual(SR) refers to the qualification rate, or selection ratio associated with a particular cutscore.

Table 17-1 was constructed using four different Taylor-Russell (1939) base rate tables (.15, .20, .25, and .30 - listed in the first column). The base rate that best fits the NF study parameters is .20 (shown in bold). That is, a .20 base rate table obtains, better than any other base rate table, the study's observed 89% success rate at the intersection of the .80 validity (the conservative population validity estimate from the study), and a .05 selection ratio (SR) (not exactly, but close). The ASVAB cutscore associated with the

.05 qualification rate, Qual(SR), is 257 ASVAB cutscore, again not exact, but close enough for a robust interpretation of expected success rates at various SRs and associated cutscores. Table 17-1 shows that the very stringent 257 cutscore, qualifying only the top 5% of the ASVAB normative population, results in an 89% success rate and that lowering the cutscore to qualify only 5% more of the population (a total of 10% at the score of 245) results in a 10% lower success rate (89% - 79%).

The obvious advantage of using the Taylor-Russell (1939) tables for cutscore analysis is that we can legitimately evaluate the expected negative impact of lowering the operational cutscore. As mentioned earlier, we cannot legitimately evaluate a cutscore lowering impact on expected success rates from the empirical cutscore analysis without making some dubious assumptions about any observable cases with ASVAB waivers (scores below the cutscore having been waivered as exception to policy without known reasons). We simply move across the SR row that is associated with specific cutscore developed from knowledge of the means and standard deviations in the PAY97 ASVAB population and find the internal table success rate values associated with the estimated population ASVAB validity. In this case, we hope for an accurate estimate of the PAY97 ASVAB composite validity from our study data. Another obvious advantage of the tables is that they are appropriate for use with a measure that is found to have higher validity than the one used operationally. In this case, one merely finds the best fit base rate table as usual for the operational composite, and then move down the validity column to the appropriate validity estimate for the candidate replacement to the expected success rate, which would always be higher or just the same if the validity increment is marginal. Of course if the recruiting environment is poor, one can establish point-waiver tolerances if lowering the operational cutscore does not hugely impact the success rate.

We can also use the Taylor-Russell (1939) tables in a reverse procedure to back out an estimate of the unrestricted validity, if for some reason we do not have the continuous performance variable to correlate with the ASVAB. For example, we might have just the dichotomous pass/fail outcome, though and we note that Chapter 12 provides a brief discussion about the correction of the range corrected validity for dichotomization. However, what if the only data available are a published empirically-based expectancy table without reference to the validity coefficient? In that case, the researcher can establish the relevant parameters from the expectancy table to find the best-fit Taylor-Russell table. That is, three parameters can be established from the expectancy table - selection ratio, success rate, and improved success rate by increasing the stringency of the selection ratio. With those three values known, the value of the validity coefficient is fixed.

Table 17-2 is used to illustrate this additional advantage of the Taylor-Russell (1939) tables. The context is the Defense Language Aptitude Battery (DLAB) for which a passing score is required for entry into the Defense Language Institute, Foreign Language Center) to study foreign languages, and the only available data is an expectancy table developed through a complex missing data procedure that addresses both multiple-hurdle and missing applicant data issue (Segall, 2007). We also present this example that applies to one Taylor-Russell table as Table 17-1 is comprised of data from four tables.

Table 17-2
Taylor-Russell (1939) .20 Base Rate Table Corresponding to a DMDC Generated DLAB Study Cutscore Table

	Selection Ratio																		
Validity	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.9	0.95
0.00	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.2	0.2	0.2
0.05	0.23	0.23	0.22	0.22	0.22	0.22	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.20	0.2	0.2	0.2
0.10	0.26	0.25	0.25	0.24	0.24	0.23	0.23	0.23	0.22	0.22	0.22	0.22	0.22	0.21	0.21	0.21	0.21	0.21	0.2
0.15	0.30	0.28	0.27	0.26	0.26	0.25	0.25	0.24	0.24	0.23	0.23	0.23	0.22	0.22	0.22	0.21	0.21	0.21	0.2
0.20	0.33	0.31	0.29	0.28	0.28	0.27	0.26	0.26	0.25	0.24	0.24	0.24	0.23	0.23	0.22	0.22	0.21	0.21	0.21
0.25	0.37	0.34	0.32	0.31	0.30	0.29	0.28	0.27	0.26	0.26	0.25	0.24	0.24	0.23	0.23	0.22	0.22	0.21	0.21
0.30	0.41	0.37	0.35	0.33	0.32	0.30	0.29	0.28	0.28	0.27	0.26	0.25	0.25	0.24	0.23	0.23	0.22	0.21	0.21
0.35	0.45	0.40	0.38	0.36	0.34	0.32	0.31	0.30	0.29	0.28	0.27	0.26	0.25	0.24	0.24	0.23	0.22	0.22	0.21
0.40	0.49	0.44	0.41	0.38	0.36	0.34	0.33	0.31	0.30	0.29	0.28	0.27	0.26	0.25	0.24	0.23	0.23	0.22	0.21
0.45	0.54	0.48	0.44	0.41	0.38	0.36	0.35	0.33	0.31	0.30	0.29	0.28	0.27	0.26	0.25	0.24	0.23	0.22	0.21
0.50	0.58	0.51	0.47	0.44	0.41	0.38	0.36	0.34	0.33	0.31	0.30	0.29	0.27	0.26	0.25	0.24	0.23	0.22	0.21
0.55	0.63	0.56	0.50	0.47	0.43	0.41	0.38	0.36	0.34	0.32	0.31	0.29	0.28	0.27	0.25	0.24	0.23	0.22	0.21
0.60	0.68	0.60	0.54	0.50	0.46	0.43	0.40	0.38	0.36	0.34	0.32	0.30	0.29	0.27	0.26	0.24	0.23	0.22	0.21
0.65	0.73	0.64	0.58	0.53	0.49	0.45	0.42	0.39	0.37	0.35	0.33	0.31	0.29	0.27	0.26	0.25	0.23	0.22	0.21
0.70	0.79	0.69	0.62	0.56	0.52	0.48	0.44	0.41	0.38	0.36	0.34	0.31	0.30	0.28	0.26	0.25	0.23	0.22	0.21
0.75	0.84	0.74	0.66	0.60	0.55	0.50	0.46	0.43	0.40	0.37	0.34	0.32	0.30	0.28	0.26	0.25	0.23	0.22	0.21
0.80	0.89	0.79	0.71	0.65	0.59	0.53	0.49	0.45	0.41	0.38	0.35	0.33	0.30	0.28	0.27	0.25	0.24	0.22	0.21
0.85	0.94	0.85	0.77	0.69	0.63	0.56	0.51	0.47	0.42	0.39	0.36	0.33	0.31	0.29	0.27	0.25	0.24	0.22	0.21
0.90	0.98	0.91	0.83	0.75	0.67	0.60	0.54	0.48	0.44	0.40	0.36	0.33	0.31	0.29	0.27	0.25	0.24	0.22	0.21
0.95	1.00	0.97	0.91	0.82	0.73	0.64	0.56	0.50	0.44	0.40	0.36	0.33	0.31	0.29	0.27	0.25	0.24	0.22	0.21
1.00	1.00	1.00	1.00	1.00	0.80	0.67	0.57	0.50	0.44	0.40	0.36	0.33	0.31	0.29	0.27	0.25	0.24	0.22	0.21

Table 17-2 is one of any number of possible mathematically generated Taylor-Russell (1939) tables (by Ms. Rebecca Hetter, formerly of Navy Personnel Research and Development Center to augment the published 10 tables). This particular "best fit" .20 base rate table shows that a validity coefficient of .55 corresponds to the pass/fail rates observed in the DMDC cutscore analysis table for four different selection ratios (5%, 10%, 20%, and 30%), which are tied to four DLAB scores (111, 102, 92, and 85, respectively), which in turn are tied to four different success rates (63%, 56%, 47%, and 41%, respectively). We intuitively know that learning a foreign language can be very difficult, especially given the time constraints that the military allows for training, and so the .20 base rate table seems appropriate. We can also see that DLIFLC has virtually no latitude in raising the already stringent DLAB cutscore of 100 (that applied to the second hardest language category) to 110 (that applied to the hardest language category) for all languages without severely limiting the DLIFLC qualified pool of applicants.

On the other hand, if not enough qualified recruits meet the 110 cutscore on the DLAB, other measures could be taken to improve the DLIFLC success rate at any cutscore level by simply (a) allocating more resources to recruiting DLIFLC candidates including screening for those who are highly motivated to learn languages, (b) increasing the time allowed to bring students to proficiency and improving other aspects of the course, or (c) adding other predictors that more closely link to the course performance and ensure that these predictors have high levels of reliability.

We note that the DLAB is a multiple-hurdle test for DLIFLC qualification in that each candidate must first meet an ASVAB standard that applies to the specific military occupation that requires foreign language training (other qualifiers include security clearance eligibility). As we know, each screen has its own selection ratio, and as we add more, the consequences are to further limit the qualified pool of applicants.

Multiple Cutscores

We discuss multiple cutscores before multiple hurdles because the multiple-cutscore topic has many interesting features, including the impact of test score unreliability on cutscore decisions. As we know, psychometrically, reliability caps validity (but reliability is not sufficient for validity) and the question becomes how unreliable was a decision to reject an individual based upon a test that has reported low reliability. Lord (1962) was concerned that perfect reliability typically is assumed on two predictor instruments each having an operational cutscore. Through extensive statistical development, Lord offered a formula that could be used to portray the expanded surface on a bivariate distribution that qualified individuals due to the unreliability of two selection instruments. Figure 17-4 is a rough portrayal of Lord's published graphic (p.28).

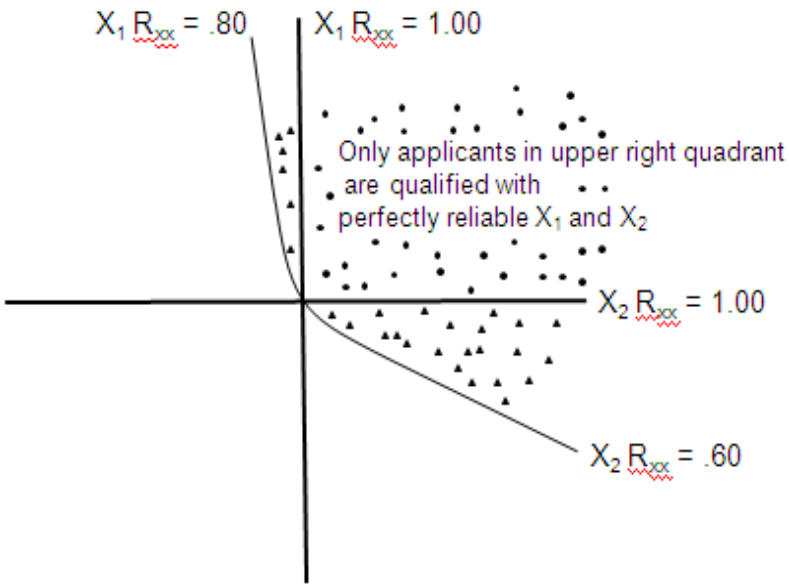


Figure 17-4. Expanded qualification surface due to measurement error.

From Figure 17-4, we see that the reliabilities of the two selection instruments will affect the quality of the individuals selected for the job as the surface thought to contain the qualified applicants (bounded by the right angle formed by the X_1, X_2 cutscores - upper right quadrant) expands to a larger coverage with the contoured line. Lord (1962) advised practitioners to consider the reliability of the instruments in setting multiple cutscores and reconsider use of the instruments if the contour line resulting from his procedures looks more like a straight line than lines at a right angle. Alternatively, if one of the two instruments is more reliable than the other, some sort of offsetting cutscore on the most reliable instrument could be considered. Another approach is to shift the contour lines upward and to the right to be more assured that the majority of those qualified would have the requisite level of the constructs. It should be noted that Lord did not consider the criterion or validity coefficient in this particular multiple cutscore focus.

Other experts have provided insights into the multiple-cutscore selection model. For example, with regards to setting effective multiple cutscores, Thorndike (1949) stated,

> "There is no really analytical way of establishing the two critical scores below which an individual shall be disqualified. Different combinations of score values for X_1 and X_2 must be tried. For each of the selected combinations the research worker must determine (1) the percentage of cases disqualified by using this combination of cutting scores and (2) the amount of difference in average criterion score for the accepted group and the rejected group. That combination of cutting scores will be chosen which (1) yields the proportion of accepted applicants which fits the supply of candidates on the one hand and the demand for job placements on the other hand and (2) makes the sharpest discrimination in criterion score between those who are accepted and those who are rejected" (p. 197).

Thorndike went on to describe that there is an advantage of multiple cutscores if one of the instruments demonstrates a dramatic non-linear relation with the criterion (e.g., where some amount of the measured construct is clearly indicated). Thorndike also suggested that a clinical approach might serve the purposes as well as an empirical approach. Cronbach (1949) cited a legitimate case for applying multiple cutscores that appears to have been determined by a clinical assessment:

> "At one time during World War II, Naval recruits were selected for training in the operation of antisubmarine listening gear on the basis of their combined scores on tests of auditory discrimination and mechanical comprehension. As a result, a number of college-trained men who excelled in mechanical comprehension but happened to be deficient in the essential auditory skills were assigned to the training, with subsequent failure. Standard Navy procedure required that those failing in the training be transferred to general sea duty as apprentice seamen. The loss of potential specialized service resulting from such a misclassification is apparent. Further analysis of the situation led in time to the substitution of a multiple-screen procedure for this selection purpose" (p. 475).

We note here that classical test theory (CTT) discussed in previous chapters tells us that we can establish with some probability the range of scores that bound a true score. However, for any individual's observed score that equals the cutscore, we cannot know if the true score is above or below that cutscore.[15]

A Navy Example of Too Many Cutscores

Table 17-3 shows a multitude of multiple ASVAB cutscores that were in place for the Nuclear Field community before they were dealt with in the late 1990s (Held, 1999) (a situation we do not want to repeat).

Table 17-3
The Nuclear Field's Prior Multiple Cutscore Qualification System

ASVAB	Cutscore	Fiscal Year 1997 recruits qualified (%)
1a VE+AR	113	30
1b VE+AR	103	60
2 AR+MK+EI+GS	218	39
3 MK+EI+GS	156	54
4 MK+AS	96	75
5 AR+2MK+GS	196	76
6 VE	41	99
All ASVAB requirements		26
1a requires NAPT	49	unknown
1b requires NAPT	55	unknown

Note. NAPT is the Navy Advanced Placement Test, a 2-hour test of advanced mathematics and physical science.

It is obvious from Table 17-3 that with so many correlated cognitive ability tests and overlapping composite components, many Navy recruits would not qualify for the Nuclear Field (NF) simply based on the many ways to be disqualified – largely based upon the compounding of test measurement error.

[15] Widely published are the formulas for deriving the confidence interval for an individual's observed score that would contain the true score with some reasonable expectation. These formulas are worked through by Harville (1991) in a web accessible paper, which is one of a series of NCME Modules dealing with educational measurement topics. www.http://ncme.org/linkservid/6606715E-1320-5CAE-6E9DDC581EE47F88/showMeta/0/

For NF, missing any cutscore – even by one point, disqualified a candidate. The solution for the NF community was to establish a two-pronged ASVAB/NAPT qualification system that addressed the issues experienced by the community: (a) not enough Navy recruits qualified for NF due to the strict adherence to the multiple cutscores, and (b) the mandatory requirement for passing the two-hour Navy Advanced Placement Test (NAPT). The administration of the test was a major qualification roadblock, not just because the test is difficult, but also because it required a Navy certified test administrator to administer the test (to over 3,000 Navy NF candidates a year).

The two-pronged classification system operationalized two ASVAB composites estimated as having the highest validity for predicting NF school grades (replacing the VE+AR composite that had the lowest validity) and allowed those that met a very high cutscore to by-pass the NAPT. We refer the reader to Appendix B of the Introductory Manual for the updated evaluation of the two-pronged NF classification system.

Figure 17-5 has been used for illustration purposes to inform Navy policy-makers at a very non-technical level about the excessive screening out of Nuclear Field candidates due to the extensive set of multiple cutscores on correlated cognitive measures.

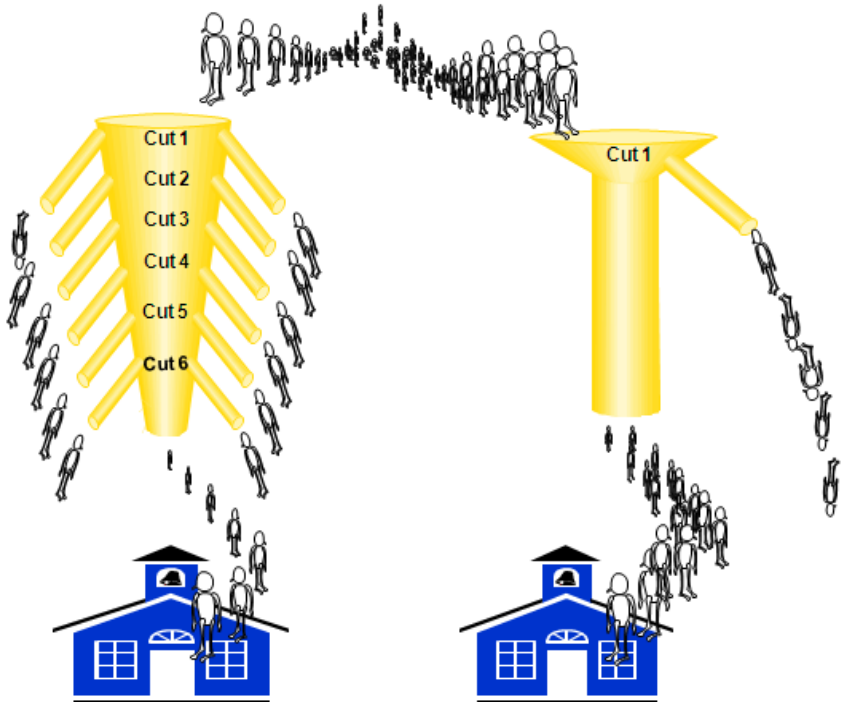


Figure 17-5. Excessive screening out due to overuse of cutscores.

Multiple Cutscores or Compensatory Model

Figure 17-6 illustrates one of the dilemmas faced when deciding whether to use a composite of measures with a single cutscore, the compensatory model, or a cutscore on each measure, the multiple cutscore model. The graph in Figure 17-6 (somewhat following Thorndike, 1982, Figure 9.1) depicts two uncorrelated selection instruments – somewhat like Figure 17-4, except in Figure 17-6 the two measures are uncorrelated and each is assumed to be perfectly reliable. To simplify matters, we assume equally weighted X_1 and X_2 variables in the compensatory model. We might consider that X_1 is a Navy Rating's ASVAB classification composite and X_2 is a personality composite that reflects, say "Agreeableness" traits (as might be useful in selecting empathic/sympathetic Hospital Corpsman). We would not expect X_1 and X_2 to correlate (many personality traits do not correlate with cognitive measures).[16] As with Figure 17-4, the criterion is not explicitly considered; however, we assume there is at least a slight increment in validity in order to consider the compensatory model to be of benefit (say, personality provides a .05 validity increment to the ASVAB's .70 for predicting training grades).

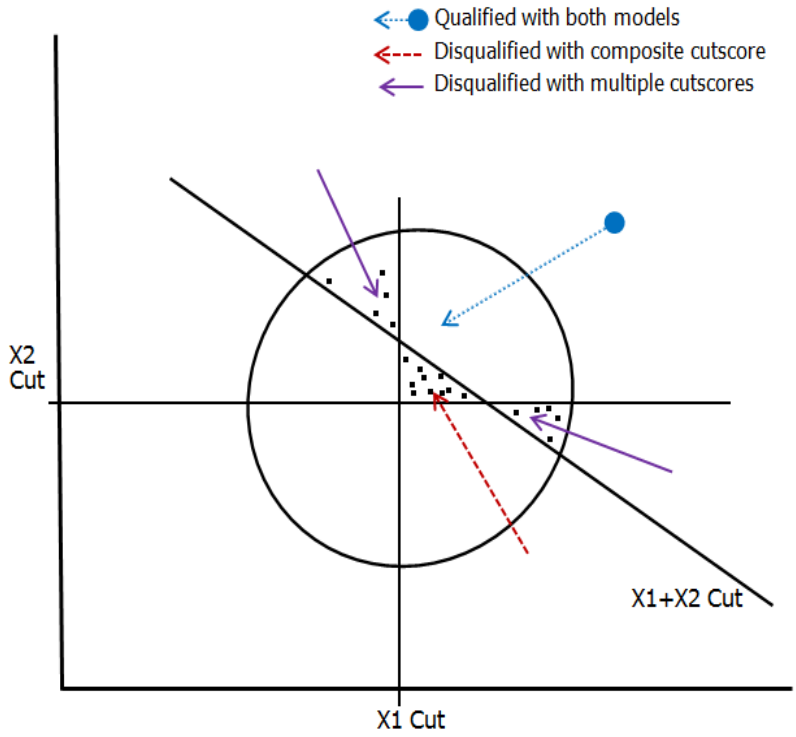


Figure 17-6. Rejected applicants from multiple cutscores vs. composite score.

[16] The more highly correlated the measures, the more overlap in qualified applicants for the two systems, all other things being equal.

Each cutscore model in Figure 17-6 qualifies the same number applicants, and therefore the same number of disqualified applicants. The diagonal X_1+X_2 composite (compensatory) cutscore score line allows higher scores on X1 to compensate for lower scores on X_2, and vice versa. The compensatory model disqualifies those in the mid-section segment of the bivariate surface (dots) who would have qualified under the multiple-cutscore model. Conversely, the X_1 and X_2 individual cutscores (like the Nuclear Field multiple cutscore model – Table 17-3) disqualifies those in the extreme segmented surfaces who would have qualified under the compensatory model. Figure 17-6 and Thorndike (1982) suggest that among the issues to consider in the comparison of a multiple cutscore and multiple-regression model is the value we place on those *disqualified* by each model.

From earlier citations from Thorndike (1949) we know that setting cutscores on a multiple cutscore model is mainly subjective, hit/miss, and iterative, but also, grounded in the position that both attributes are required for the job. Add to this complex web of considerations for the military: (a) not dealing with just one occupation, but a system of many occupations, (b) changing recruiting environments that can dramatically change the supply of applicants with high levels of one or the other attribute (the military is most concerned about ASVAB score distributions and education), and (c) personality may really matter most for job performance whereas the ASVAB may matter most for training performance (upon which ASVAB is currently validated).

We note that the operational qualification model is multiple cutscore for both the NAPT (for the Nuclear Field Ratings) and the DLAB (for foreign language required military occupations), but that these instruments are only administered to those qualifying on occupational ASVAB cutscores, and thus follow a multiple-hurdle model.

Cutscores for Multiple Hurdles

As discussed in Chapter 16, with multiple-hurdle selection systems, applicants must pass one screen to proceed to the next and so forth until all of the hurdle screens have been passed. For simplicity, we do not discuss the more expensive second-stage hurdles such as personnel interviews, background security checks, and assessment centers, but only measures that are administered either on the CAT-ASVAB platform or by other testing means (e.g., by test control officers or their test administrators in the case of the NAPT). The multiple-hurdle model is efficient to both the organization and the individual (see Chapter 15). However, just as with multiple cutscores, multiple-hurdle cutscores are complicated. Hunter, Schmidt, and Le (2006) considered the multiple-hurdle topic important but complicated and pointed out the many possible cutscores that could be assigned to each hurdle instrument, depending upon the order and validity magnitude of each. (Because the ASVAB is a given to all military applicants, it makes sense that it should be the first selection/classification hurdle.)

Sackett and Roth (1996) noted that the ultimate selection ratio resulting from two hurdle cutscores is the product of the two separate hurdle selection ratios. We consider that this principle applies to instruments that are uncorrelated such as the near zero (or zero) correlation of the ASVAB with many facets of a personality instrument. The ultimate (or net) selection ratio principle being the product of the two is obvious when we consider that contributions to the multiple correlation for uncorrelated measures in regression analysis are additive. We also note that with correlated measures (such as the multiple cutscore model formerly used by Nuclear Field, the reduction in the selection ratio is not so severe and is ascertained by merely cutting the recruit data on all ASVAB requirements (as all recruits have ASVAB scores). The task is more complicated when a second hurdle instrument is correlated with the ASVAB (e.g., NAPT and DLAB) and not all recruits have been administered the test (Segall, 2007, used Monte Carlo Markov Chain to estimate DLAB scores for an applicant population).

For example, consider the situation where a personality measure (say achievement traits) has a known small validity coefficient for predicting training performance, but zero incremental validity to the ASVAB. Because of personality's small validity coefficient (say r_{xy} = .20), a very lenient cutscore is set—one that qualifies 80% of the applicant population. Now, say, as in the Navy's Nuclear Field (NF) situation, the predictive validity of the ASVAB is large (r_{xy} is estimated at about .80 - .85 in the Appendix B study of the Introductory Manual). For NF, the training is extremely difficult, and there is a larger proportion of failures historically observed when the ASVAB cutscore is set even only slightly lower (say qualifying the top 10% of youth rather than the current top 5%). The effective selection ratio according to Sackett and Roth (1996) using the ASVAB cutscore qualifying the top 10% of youth and a cutscore on the personality instrument qualifying the top 80% is .80 X .10 = .08 for a lower 8% qualification rate. A 2% loss in qualified recruits seems small until you consider that (a) 2,500 NF candidates are required out of less than 40,000 Navy accessions each year and (b) many other Ratings also require high ASVAB-scoring youth. Of the 2,500 NF annual recruiting goal, a two-percentage-point reduction due to the "personality hurdle" would result in 800 fewer recruits qualified for NF out of the 40,000 accession population.

Establishing a cutscore for an instrument used in a multiple-hurdle selection system in predicting performance in Navy training would require evidence that the instrument (a) had a direct linkage to the underlying performance constructs and provided stability of measurement (i.e., was not an unreliable measure), (b) had a relation with the training measure that was confirmed over time (e.g., not subject to sample fluctuations and statistical significance by chance), (c) was a measure of a relevant construct not measured by the ASVAB or if measured, not measured well (such as extremely high aptitude/ability levels as was the case with the NF Ratings, (d) does not inordinately reduce the ability of the Navy to fill the occupation (as was the case for the NF with the strict requirement of the NAPT hurdle), and (e) is well monitored to protect against test item security issues (i.e., items put on the internet) and test compromise (cheating in various ways).

Managing ASVAB Cutscore Waivers

During a poor military recruiting environment (e.g., from a positive U.S. economy with an abundance of good jobs), it becomes more difficult for the military to recruit a large number of very smart youth. It may become necessary to issue ASVAB score waivers, if not military qualified, depending upon the Service, and also for certain complex/high aptitude/ability military occupations. Without significant offsets such as high bonuses, fewer ASVAB scores will be observed in the upper tail of applicant and recruit ASVAB score distributions. In this case, we must be cautious about setting cutscores or advising on cutscore waiver score points using the Taylor-Russell (1939) tables because we may not meet the underlying bivariate normal distribution (ASVAB and performance distributions). Smith (1948) discusses the issue of not meeting distribution assumption in use of the tables.

At this time, it is unclear what policy will be issued to Navy Recruiting Command for allowing ASVAB score point waivers for Navy Rating classification. The current prolonged recruiting environment will not last forever (at least if history repeats itself), and there will be tensions between recruiting and training about filling goal and filling Fleet requirement with adequately trained Sailors. We note that a 12-point waiver (3-points per ASVAB test or about, 1/3 of a standard deviation – the largest allowed waiver applying to four-test composites) does not have the same negative impact across Ratings, all other things being equal including a substantial course fail rate.

We have two anecdotal stories to tell about lowering ASVAB standards, which is in essence what is done when an ASVAB point-waiver is issued. The first applies to the Mineman (MN) community when it was allowing Sailor conversions into its Rating during a period when the Navy required growth in the community to address a long-term mission spike for minesweeping capabilities (detection and avoidance of mines at sea). Some Sailors were allowed to convert to the MN Rating from low-tech occupations. As a result, ASVAB point-waivers were given, the Sailors shipped to shore-based training, and such extreme learning issues resulted that NPRST was asked to "raise the ASVAB standard" and provide a separate Electronics Information (the ASVAB EI test) cutscore. The study is provided in Appendix C.

The other anecdote occurred during the time that the ASVAB Coding Speed (CS) test was eliminated from the battery (in 2002 when the ASVAB Assembling Objects test was added). The Navy, the only Service that supported CS, was able to retain it for classification testing (on the CAT-ASVAB platform) because it demonstrated that the test (a) offers incremental validity to otherwise optimally-formed ASVAB composites for a substantial number of Ratings, (b) reduces adverse impact because it is a relatively culture-free test, and (c) increases the proportion of recruits in any given year who are qualified in the aggregate across military occupations. Eliminating CS would have negative impacts in all three areas.

The Marine Corps addressed the negative impact of CS's removal from the ASVAB in the ability to fill all of their occupations by lowering their ASVAB classification composite cutscores by 5-score points across all occupations. (The Marine Corps' ASVAB composite scores are standardized to have means of 100 and standard

deviations of 20, so this amounted to a .25 standard score lowering of the cutscore.) Naturally, an across-the-board lowering of the qualification cutscore would be expected to have differential performance impacts depending upon the training parameters (difficult, time constraints, observed academic failure rate) and ASVAB validity coefficient. These factors were not taken into consideration. The Navy and Marine Corps train together for the occupation of Air Traffic Controller (ATC). As it happened, the Director of Training at the schoolhouse called for an ASVAB validation/standards review shortly after the elimination of CS from the ASVAB to better understand the jump in academically-related failure rates in the schoolhouse. As it turned out, a visit to the schoolhouse revealed that the uptick was due to the lowering of the Marine Corps ATC ASVAB standard.

Guidance for Issuing ASVAB Cutscore Waivers

Another complicating factor involving ASVAB cutscore point-waivers is that many of the Navy Ratings have alternative ASVAB standards. For example, candidates for the three Nuclear Field Ratings and the Mineman Ratings (see Appendices B and C of the Introductory Manual) can qualify on the VE+AR+MK+MC or AR+MK+EI+GS composites (Mineman Rating having a lower cutscore than the Nuclear Field Ratings). Although a model permitting alternate ASVAB standards has definite merits in opening the aperture for occupational qualification, the model also increases the potential for applicants to qualify on chance factors. Just as the Nuclear Field initial multiple cutscore model excluded many applicants based upon measurement error, qualifying on alternative standards includes some applicants based on that same measurement error. It is easy to see that a candidate who only qualifies on one of the two standards by just one score point may be only marginally qualified; others should be considered better choices. Unfortunately, during recruiting downturns, the military cannot be assured that waiting for the more qualified candidates will result in meeting a fixed yearly goal for highly technical occupations.

The following factors should be considered when recommending leniency of an ASVAB cutscore waiver policy for any Navy Rating involved in an ASVAB validation/ standards study:

1. alternative ASVAB standards or single ASVAB standard;
2. cost of training resources wasted when students fail (highly dependent on course length);
3. training methods and remediation capacity that would allow students having difficulty to recycle;
4. criticality of the Rating to Fleet operations where at-the-margin students might cause accidents or risks to other Sailors;
5. yearly input requirement, which if a large number, could produce stress in filling the Rating; and
6. a changing recruiting environment.

Concluding Remarks

This chapter provided a brief background of various methods and models for setting cutscores, some of which apply directly to the Navy. All Navy Ratings have an ASVAB standard and so we are not at square one in setting cutscores, unless a new Rating is stood up. Even at that, we can obtain a lot of information about that Rating and its proposed training and apply an initial cutscore using an ASVAB normative process. Cutscores for the military are set in a fluid environment as resources wax and wane over time, leading the ASVAB validation/standards program to be considered a necessary operational maintenance requirement. This requirement will become more evident as more joint-service training occurs and the varying ASVAB cutscores come into question. Both the Army and Navy have simulation software applications that incorporate cutscores for assessing the requirement for ASVAB waivers. Two of the Navy applications are described in the next chapter, one of which allows, for any given Navy Rating, the designation of ASVAB points and the percentage of allowed waivers.

Chapter 17. References

Aguinis, H. (Ed.) (2004). *Test-score banding in human resource selection: Technical, legal, and societal issues*. Westport, CT: Praeger Publishing Company.

Angoff, W.H. (1971). Scales, norms and equivalent scores. *Educational Measurements*. Washington, DC: American Council on Education.

Cascio, W. F., & Aguinis, H. (2005). Test development and use: New twists on old questions. *Human Resource Management, 44,* 219-235.

Cronbach, L. J. (1949). *Essentials of psychological testing*. NY: Harper.

Cureton, E. E. (1957). Recipe for a cookbook. *Psychological Bulletin, 54,* 494-497.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. San Francisco, CA: W. H. Freeman and Company.

Harville, L. M. (1991). An NCME instructional module on standard error of measurement. *Educational Measurement: Issues and Practice, 10,* 33-41.

Held, J. D. (1999). *Eliminating multiple cutscore aptitude qualification standards for Navy enlisted training*. Proceedings of the 41st Annual Conference of the International Military Testing Association, pp. 159-164. Monterey, CA.

Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology, 91,* 594-612.

Lord, F. M. (1962). Cutting scores and errors of measurement. *Psychometrika, 27,* 19-30.

Sackett, P. R., & Roth, L. (1991). A monte carlo examination of banding and rank order methods of test score use in personnel selection. *Human Performance, 4,* 279-295.

Sackett, P. R., & Roth, L. (1996). Multi-stage selection strategies: A monte carlo investigation of effects on performance and minority hiring. *Personnel Psychology, 49,* 549-572.

Segall, D. O. (2007). *The Utility of the Armed Forces Qualifying Test as a Predictor of Language Aptitude and Proficiency*. Information paper. Seaside, CA: Defense Manpower Data Center.

Segall, D. O. (2004). *Development and evaluation of the 1997 ASVAB score scale* (Technical Report No. 2004-002). Seaside, CA: Defense Manpower Data Center.

Smith, M. (1948). Cautions concerning the use of the Taylor-Russell tables in employee selection. *Journal of Applied Psychology, 32,* 595-600.

Taylor, H. C., & Russell, J. T. (1939). The relation of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology, 23,* 565-578.

Thorndike, R. L. (1949). *Personnel selection*. NY: Wiley.

Thorndike, R. L. (1982). *Applied Psychometrics*. Boston: Houghton Mifflin.

Truxillo, D. M., Donahue, L. M., & Sulzer, J. L. (1996). Setting cutoff scores for personnel selection tests: Issue, illustrations, and recommendations. *Human Performance, 9,* 275-295.

Chapter 18.
Assessing ASVAB Standards Adequacy through Simulation
Janet D. Held

Introduction

As discussed in the last chapter, setting cutscores is a complicated matter, even for one job. The process is even more complicated for the military when there are many jobs and variations in the recruiting environment. As one thread of ASVAB validation/ standards research projects, the Navy partnered with industry in the development of two job classification/assignment applications to help assess the impact of ASVAB standards changes on the Navy's ability to fill all Navy Ratings. This chapter describes the two applications and their research use in the evaluation of the classification effectiveness of two tests, Assembling Objects, which was added to the ASVAB in 2002, and Coding Speed, which was eliminated from the ASVAB at the same time.

The Focus on Opening the Aperture for Occupational Qualification

There is an extensive literature on adverse impact (e.g., Sackett, Laczo, & Lippe, 2003) (with additional context and references cited in Chapter 13), including the concept and how various terms like "fairness" evolved (e.g., Lawshe, 1987). The Navy recognizes that the ASVAB technical tests present occupational barriers (adverse impact) to women and some minority groups that are not familiar with the technical knowledge-based test content (e.g., Electronics Information and Auto/Shop Information). However, when combined with more academic tests, these tests are highly valid in predicting performance in a wide range of technical Ratings and the thought is not to penalize those with high technical test scores by removing the tests from the battery. The Defense Manpower Data Center (DMDC) in the early 1990s led a joint-service evaluation of the fairness of the ASVAB technical composites for technically oriented jobs (Wise et.al, 1992). The study concluded that although the ASVAB was *not* biased in predicting training success for minority groups (equal regression weights for majority and minority groups), but that some effort should be made to reduce ASVAB score barriers (adverse impact). Wise et al. recommended that score barriers be addressed by adding a fluid intelligence test (Carroll, 1993; Cattell, 1943) from the Enhanced Computer-Administered Test (ECAT) battery (Alderton, Wolfe, & Larson, 1997) to complement the ASVAB tests, which were mostly measures of crystallized intelligence at the time.

The Navy responded to the Wise et al. (1992) recommendation by conducting further ASVAB predictive validity studies involving Assembling Objects (AO) (Held, Fedak, Crookenden, & Blanco, 2002; Held, Fedak, & Johns, 2004). The validation studies resulted in the inclusion of the AO test in two Navy ASVAB composites (see Table 6-1 in the Introductory Manual). Primarily, the Navy uses AO in ASVAB composites used to classify recruits into mechanical (and engineering) oriented Ratings, which is consistent with prior studies on AO (e.g., Carey, 1994).

AO was officially added to the ASVAB at the same time that the two ASVAB speeded tests, Numerical Operations (NO) and Coding Speed (CS), were removed. Because NO and CS load on a speeded ASVAB factor, their elimination increased the average intercorrelation of the remaining ASVAB tests, thus reducing the potential to differentially assign recruits to occupations for which they are best aptitude/ability suited. The average ASVAB intercorrelation of the battery's tests for the Profile of American Youth 1980 (PAY80) study was .593 when the ASVAB included the NO and CS test (the PAY80 matrix was taken from Maier & Sims, 1986). Removing NO and CS increased the average test intercorrelation to .655 (the AO test was not a part of the ASVAB for this intercorrelation assessment).

The Navy provided evidence through the Manpower Accession Policy Working Group (MAPWG) to support the continued inclusion of CS in the ASVAB, but evidence was not sufficient for the NO test. Inclusion of the CS test (a) provided incremental validity to the existing ASVAB composites when validated against training performance for relevant Navy Ratings (Held & Wolfe, 1997); (b) lowered the average intercorrelation of the ASVAB tests in a large Navy recruit population, implying increased differential assignment capability; and (c) lowered score barriers for women and some racial/ethnic minority groups (Alderton et al., 1997; Held et al., 2002).

Although the Navy found clear support for the CS test, the problems of maintaining the tests were considered non-trivial for the long term, and so CS along with NO were eliminated from the ASVAB (Segall, 1997). However, because the Navy had sufficient evidence to retain CS, the Navy was able to retain the test as a Navy special test with administration to Navy applicants immediately (seamlessly) after completion of the computerized version of the ASVAB (CAT-ASVAB). The latest hardware effects study of the CS test shows no score differences between groups that input CS answers via the specially configured CAT-ASVAB keyboard, and a mouse so the test is considered stable (Pommerich, 2013). There also were efforts underway to further eliminate hardware effects so that the test could be administered on other computer platforms (e.g., internet-based proctored CAT-ASVAB) (Segall, 2010). The Navy assumes that at some point, the Army will be favorable to re-using the CS test, as they developed it (Held & Carretta, 2013). At least two Army reports support use of CS (Scholarios, Johnson, & Zeidner, 1994 in the capability to increase differential assignment capability; Zeidner, Johnson, Vladimirsky, & Weldon, 2004 for increasing the slopes of minor group regression lines).

As part of the evidence-gathering for lowering ASVAB score barriers, NPRST calculated ASVAB test effect sizes for gender and some clearly defined racial/ethnic groups having adequate sample sizes. The effect sizes were calculated as the majority group mean minus the minority group mean, the result then divided by the groups' population standard deviation. Table 18-1 (taken from Held & Carretta, 2013, originating from Held et al., 2002), shows the effect sizes (positive values favoring the major group) from a year 2000 recruit population.

Table 18-1
ASVAB Effect Sizes for a Year 2000 Navy Recruit Population

	Male Caucasian Reference Group N = 22,230				Female Caucasian Reference Group N = 4,454			
ASVAB/*n*	Af. Am. 6,117	Hisp. 4,049	Asian 1,777	Native Am. 1,523	Af. Am. 1,911	Hisp. 1,005	Asian 383	Native Am. 410
GS	0.93	0.68	0.78	0.03	0.87	0.68	0.53	0.16
AR	0.70	0.31	0.09	0.03	0.62	0.29	-0.07	0.10
VE	0.65	0.59	0.73	-0.01	0.66	0.57	0.45	0.12
MK	0.19	0.04	-0.42	0.05	0.11	0.02	-0.41	0.06
MC	0.93	0.43	0.43	-0.01	0.83	0.42	0.34	-0.03
AS	1.13	0.73	1.04	-0.11	1.09	0.84	0.94	0.01
EI	0.76	0.52	0.46	-0.01	0.68	0.61	0.39	0.14
AO	0.58	0.18	-0.04	-0.05	0.58	0.22	0.02	-0.03
CS	0.21	0.10	-0.08	0.06	0.17	0.18	-0.10	0.07

Note. Effect sizes were computed as the Caucasian (majority group) mean minus the minority group mean divided by the pooled standard deviation for both groups. VE is a weighted composite of the Word Knowledge (WK) and Paragraph Comprehension (PC) tests. (See Chapter 2 of the Introductory Manual for ASVAB test descriptions.)

Table 18-1 shows that, for both men and women, similar effect size patterns emerged for the racial/ethnic groups, suggesting cultural differences. Only effect sizes greater than 0.5 are discussed here, as Cohen (1988) considers half a standard deviation (SD) difference in group means to be a medium effect size (0.2 small, 0.5 medium, and 0.8 large). Table 18-1 shows the largest ASVAB test effect sizes were for African Americans, followed by Hispanics and Asians. No substantial effect sizes using this 0.5 SD criterion were found for Native Americans for either males or females, so comparisons involving the Native American group were not considered further. That said, Auto and Shop (AS) had the largest effect size (favoring Caucasians) for all majority and minority group comparisons. Mathematics Knowledge (MK) had the only negative effect size, which favored Asians over Caucasians but was a little below the 0.5 "moderate effect" threshold. Coding Speed (CS) had the lowest effect sizes across all race/ethnicity groups and within gender. Assembling Objects (AO) had a meaningful 0.5 effect size for only African Americans, but this effect size was much lower than for, the technical knowledge subtests (e.g., AS was slightly above 1.0 for both men and women).

We note that DMDC tracks ASVAB subtest effect sizes over time for gender and ethnic groups and that item screening processes are in place to address bias and adverse impact. The Navy considers both validity and adverse impact, however, not with test weighting schemes, but by offering alternative standards for all applicants that are equally valid (if not more with use of the neutral test, like CS).

The data used to develop racial/ethnic effect sizes within gender reported in Table 18-1 also were used to develop ASVAB gender effect sizes. Figure 18-1 graphically shows the gender effect sizes, again using the groups' pooled standard deviation in the denominator and the male minus female mean score difference in the numerator.

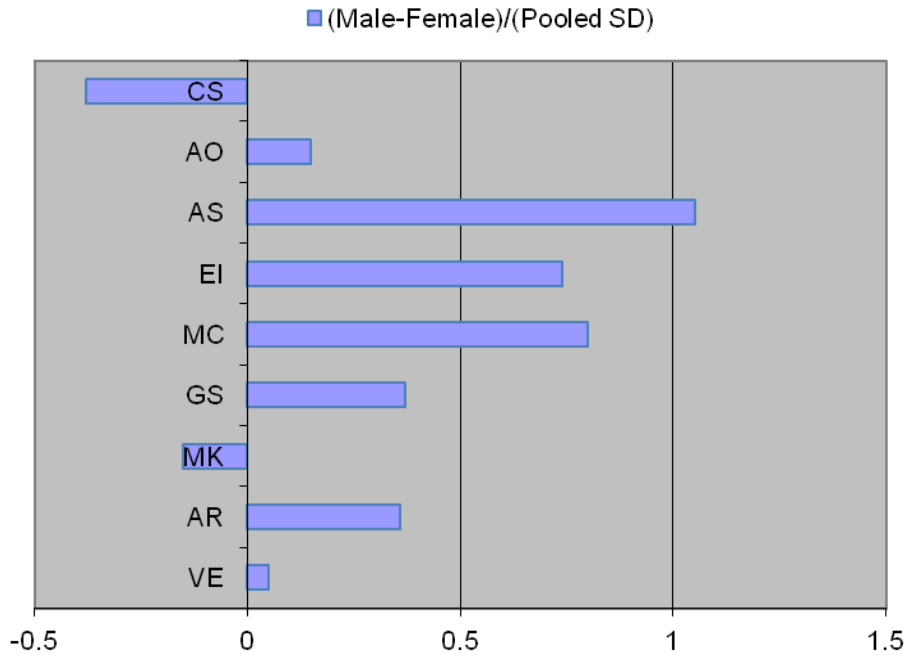


Figure 18-1. Male-Female ASVAB effect sizes for a year 2000 Navy recruit population.

Figure 18-1 clearly shows that for gender, the largest ASVAB test effect size (slightly over 1.0) was for the ASVAB Auto and Shop Information test (AS) favoring males. The AS test also showed the largest effect size favoring Caucasians in the race/ethnic analysis (Table 18-1). The effect sizes for the two other ASVAB technical tests, Mechanical Comprehension (MC) and Electronics Information (EI) were above the 0.5 threshold (between 0.5 and 1.0). The only tests that favored females, but not reaching the 0.5 threshold (with negative signs), were Coding Speed (CS) and Mathematics Knowledge (MK). The CS test, which measures processing speed and accuracy, in a detailed oriented clerical context, is consistent with the literature (e.g., Majeres, 1988).

We make a point at not disparaging the ASVAB technical tests for their gender effect size differences because (a) those who perform high on these knowledge-based tests may gravitate to military service where they know that many of the occupations are more technical than academic and (b) the incremental validities documented for the technical tests are substantial when added to the ASVAB academic tests for many military occupations, thus substantiating the tests' utility.

Although the Navy provided adequate evidence for supporting the use of CS (and later, AO), there was not sufficient evidence that differential assignment capability would be negatively impacted if the CS test were eliminated from the ASVAB. That is, just because we could point to a lower average intercorrelation of the ASVAB tests minus CS, we could not know how, operationally, the Services would be impacted in filling all of their jobs on newly configured ASVAB classification composites that did not contain the CS test. Thus, the Navy took further steps to illustrate the potential negative classification impact by applying two simple classification algorithms (in contrast to its then complicated utility index that integrated six utility functions, such as attrition and job complexity).[17] Two Navy-sponsored applications that incorporated the simple algorithms and their application to the CS and AO tests are described in the two following sections.

The Navy's Selection and Classification of Recruits Evaluator (SCORE)

When it became apparent that the ASVAB speeded tests, CS and NO, were to be eliminated in 2002, the Navy (and other Services) became concerned about the impact on being able to differentially assign recruits to their best-fitting jobs. Two applications were developed for this use, sponsored by the Navy's Selection and Classification office. The first application to be developed was called the Selection and Classification of Recruits Evaluator (SCORE), a stand-alone application incorporating the Navy's new and current classification algorithm, Rating Identification Engine (RIDE; Crookenden & Blanco, 2002; EDS Federal, 2001; Watson, 2004, 2010). RIDE/SCORE incorporates a sequential assignment algorithm that ranks jobs for each individual (randomly drawn) based on, mainly, two utility functions, each of which uses ASVAB data. The first utility function is developed from logistic regression with the Rating-specific ASVAB composite predicting first-pass pipeline success (no academic failures or academic setback incidents).

This first SCORE utility function (utility functions developed specifically for this application, not to be confused with the Brogden-Cronbach-Gleser utility model discussed in Chapter 3) adds points for individuals who have high ASVAB composite scores. The second utility function is a counterbalance and subtracts points (based upon AFQT, see Chapter 2 of the Introductory Manual) for individuals who are clearly overqualified for a particular job and therefore would not be optimally challenged. We note that the RIDE algorithm is more complex than just stated and includes several constraints; however, NPRST's use of the SCORE application that incorporates some of the functionality had utility for specifically studying the narrow impact of ASVAB changes on Navy Rating fill. (For both models/applications described in this chapter, the assignment output files can be used to predict training performance from ASVAB scores.)

[17] RIDE (Rating Identification Engine) replaced CLASP (Classification and Assignment within PRIDE) for the Navy in 2011. Most publications comparing the Services' classification/assignment systems refer to CLASP, so we provide several references (Kroeker & Folchi; 1984; Kroeker & Rafacz, 1983).

For the simulations, four sets of composites were formed for assigning recruits (input file) across Navy Ratings (sometimes referred to as jobs in the table because assignments were actually made to different programs within Ratings). Composite Set 1, the baseline set, did not apply AO or CS in any of its composites. Composite Set 2 included two composites with AO, but none with CS. Composite Set 3 included two composites with CS, but none with AO. Composite Set 4 included both the AO and CS composites formed from Composite Sets 2 and 3. All composites that contained AO and CS had been validated in previous work as adding incremental validity to the ASVAB. None of the composites contained both AO and CS. Over 150 Program/Rating ("jobs") were involved in the simulation analyses.

Improved classification from the simulated assignment using composites with AO and CS was defined in two ways: (a) an increase in the percentage of the recruit population classified to jobs in the aggregate and (b) a decrease in the standard deviation (SD) of the fill rate among different types of Ratings (stability in assignments to jobs). Separate simulations were performed for males and females because some Navy jobs are closed to females. Table 18-2 lists both male and female results.

Table 18-2
SCORE Classification Simulation Results

	Composite Set			
	Set 1 (neither AO or CS)	Set 2 (with AO)	Set 3 (with CS)	Set 4 (with AO and CS)
Scenario #1: 1.7% less female jobs than females (8,134 jobs; 8,275 females)				
Unassigned Recruits	469	413	389	288–303 (range with 4 runs)
Job Fill Standard Dev.	16.1%	15.1%	14.6%	13.7–14.8%
Scenario #2: 2.5% more female jobs than females (8,484 jobs; 8,275 females)				
Unassigned Recruits	501	440	279	279–300 (range with 4 runs)
Job Fill Standard Dev.	20.2%	18.7%	16.7%	16.4–17.4%
Scenario #3: 6.2% more male jobs than males (38,402 jobs; 36,154 males)				
Unassigned Recruits	938	661	785	492–555 (range with 4 runs)
Job Fill Standard Dev.	13.8%	13.4%	13.0 %	12.6–14.4%
Scenario #4: 13.4% more male jobs than males (40,995 jobs; 36,154 males)				
Unassigned Recruits	387	71	213	0 for all 4 runs (range with 4 runs)
Job Fill Standard Dev.	15.9%	15.6%	15.6%	18.2–19.3%

Table 18-2 shows the four composite sets listed across the table header row: (a) Composite Set without AO or CS (baseline); (b) Composite Set with AO; (c) Composite Set with CS; and (d) Composite Set with AO and CS. Table 18-2 also shows two scenarios for females and two for males and the simulation results listed in blocked rows. For females, Scenario #1 specified slightly less female jobs (1.7%) than females available to fill them, and Scenario #2 increased that percent a bit (2.5% more female jobs than females available to fill them). For males, Scenario #3 specified substantially more male jobs (6.2%) than males available to fill them. Scenario #4 increased the stress on recruiting specifying 13.4% more male jobs than males available to fill them. For the stability of assignment outcomes, the index reported was the standard deviation of fill rate (across all jobs). Scenario #4 involved four assignment simulation runs, but only for the Composite Set that included both AO and CS and so the table shows the range of standard deviation values.

Table 18-2 shows that, for every classification simulation scenario, providing an ASVAB composite set that included some composites with the AO or CS tests and especially with both where the benefits are additive, resulted in fewer recruits "unassigned" to jobs. We highlight that although both male and female assignments were increased with just the addition of AO; the benefit of AO was largest for Scenario #4, the stressed recruiting environment scenario. That is, in Scenario #4, the baseline composite set *without* either AO or CS left 387 males unassigned. By adding AO to some composites (where validity warranted AO use) there was a substantial (316) reduction in the number of unassigned males (to 71), which shows that AO has high utility for filling jobs. There was also a female assignment improvement with the AO test (Composite Set 2) but the improvements were not as great as for males most likely because many of the male types of mechanical/engineering jobs that were billeted on ships were not open for females during the study timeframe.

Finally, the standard deviation of the fill of jobs (indicating evenness of the distributed assignments) tended to decrease going from the baseline composite set to the composite set that included both AO and CS. The obvious exception was for Scenario #4 (13.4% more male jobs than males available to assign) where *everyone* was assigned to a job. Obviously with more "job choices", there was a less even fill as there was more flexibility in who got assigned to what job (standard deviations ranged from 18.2% to 19.4% for the four simulations compared to about 15%-16% for the three other composite set single runs). There is more good news than bad with the larger standard deviations under the stressed recruiting scenario as AO and CS clearly demonstrate enhanced differential assignment capability.

The Navy's Selection and Classification Cost Effectiveness Model (SCCEM)

The Selection and Classification Cost Effectiveness Model (SCCEM) (Hogan & Simonson, 2004a, 2004b) is simple in concept but with some useful functional features. As with SCORE, the SCCEM incorporates a sequential assignment algorithm, and also similar to SCORE, considers only the Navy Ratings' ASVAB standards (composites with cutscores) and each recruit's ASVAB scores (for ASVAB validation researcher use).

The SCCEM algorithm matches every Navy Rating's ASVAB standard (composite with cutscores) to that of a randomly drawn recruit (from a data file containing data for a recruit population) and assigns that recruit to the Rating that yields the lowest ASVAB score difference between recruit and Rating. (Rating ties are broken with a random assignment subroutine.) The SCCEM's "just barely qualified" algorithm, a bottom-up fill strategy, results in adequate ASVAB distributions across Ratings, and not just everyone at the margin of all Ratings' ASVAB cutscores, because small delta scores are used up in the early assignment stages (remembering the ASVAB, like most test batteries, exhibits somewhat normal test score distributions properties in unrestricted populations).

The SCCEM application has two adjustment features that allow the user to (a) input ASVAB score point waiver tolerances for any specific Rating and the proportion of recruits that can have a waiver or (b) alternatively, fixed input waiver parameters that are applied across all Ratings. A Rating can be split into two smaller Ratings (e.g., NFa and NFb) with half of their annual recruiting goal numbers assigned to each to account for the use of alternative ASVAB standards (e.g., VE+AR+MK+MC or AR+MK+EI+GS in the case of Nuclear Field).

Finally, the application calculates recruiting costs for the required "quality" across Ratings where quality is defined by high school diploma status and AFQT score. Figure 18-2 shows a DoD matrix that defines the "quality" cell combining AFQT and Education.

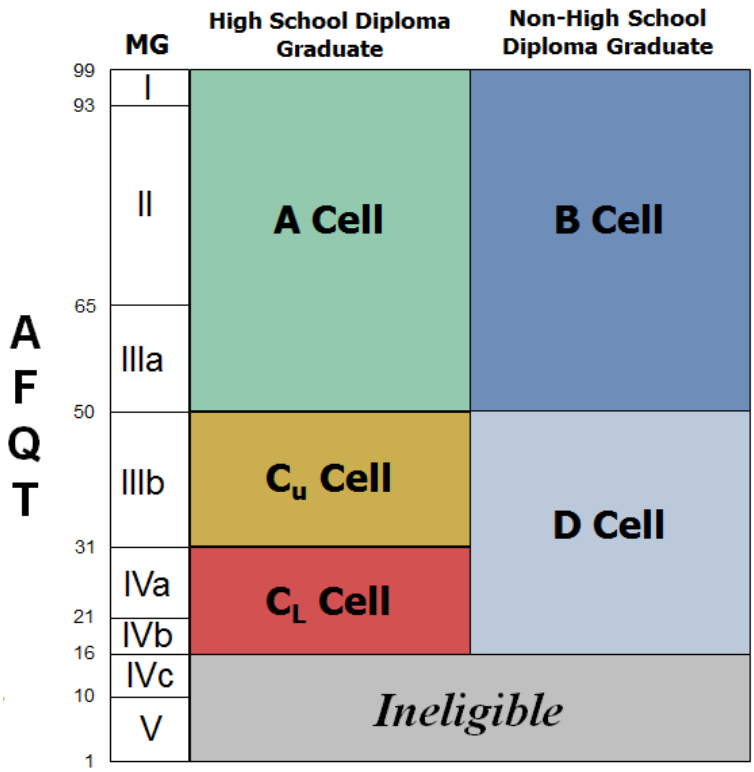


Figure 18-2. Recruit ASVAB and education quality cells.

Figure 18-2 shows that an AFQT score of 50 or above and a high school diploma graduate (HSDG) are considered A-Cells. A-Cells are the most expensive to recruit because these individuals generally have other than military options to choose from, such as high-paying jobs or college opportunities. Those with AFQT scores of 50 or above but without HSDGs (B-Cell) are the second most expensive to recruit, but the number of non-HSDGs is constrained due to DoD policy.[18] Those with AFQT scores from 31 to 49 and a HSDG (C-Cell) are the least expensive to recruit, at least during the period of the SCCEM's development.

An additional functionality resulting from the SCCEM application, and also SCORE, as we have seen, is the ability to study the potential for diversity in race/ethnicity/ gender spread across Ratings from the Rating assignment simulations from adding tests such as AO and CS to the ASVAB. Indirectly, lowering adverse impact by adding these tests lowers recruiting costs because the AO and CS tests are less correlated with the AFQT (CS less correlated than AO) that is tagged to recruiting costs. To illustrate the point, Figure 18-3 displays the results of one SCCEM simulation applying the ASVAB Composite Set 1 (baseline without AO or CS) from the SCORE study and Composite Set 4 (including both tests).

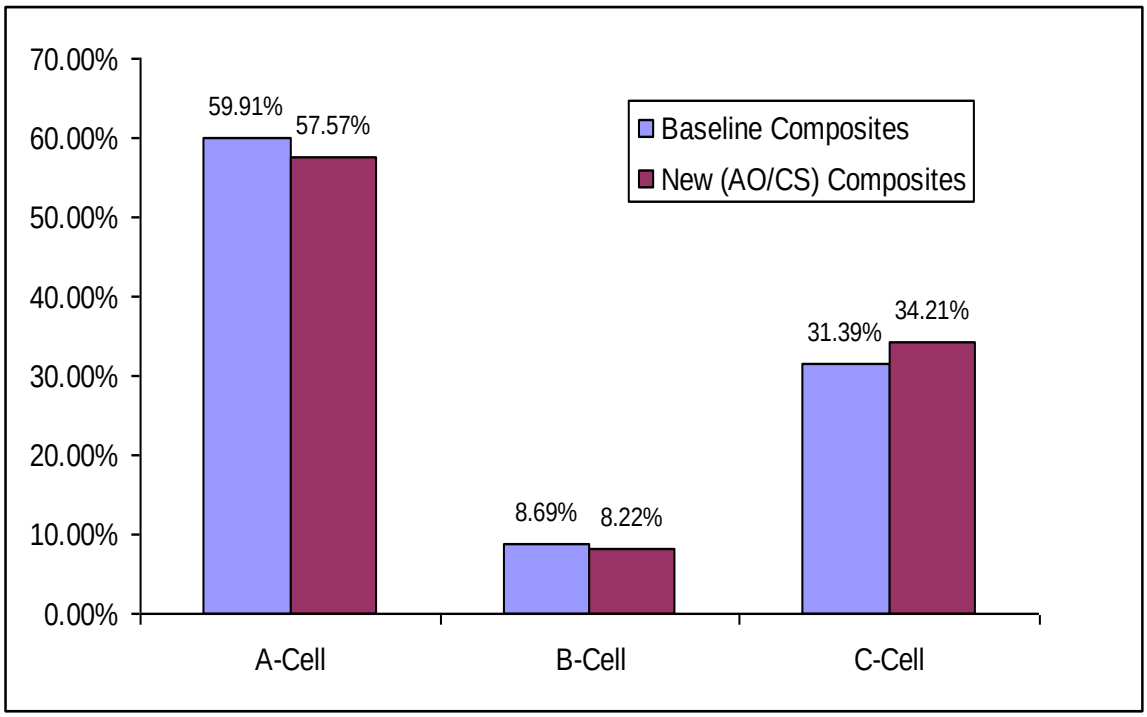


Figure 18-3. SCCEM simulation results (from Hogan & Simonson, 2004b).

[18] The National Defense Authorization Act, 2012 (NDAA2012) included some previous educational credentials educational that were not considered A-Cell high school diploma graduates in a revised "Tier" system. In this revision, for all intents and purposes, some previous Tier II classifications are now Tier II without limits on recruitment.

Figure 18-3 shows that using AO and CS in ASVAB composites resulted in some less expensive C-Cells filling Navy Ratings that otherwise would have to have been filled with more expensive A-Cells. Although this was not the case for B-Cells, B-Cells become irrelevant because there are relatively few of them recruited (at the time).

Hogan and Simonson (2004) adopted a conservative cost assumption for the simulation: $12,000 for A-Cells, $4,000 for B-Cells, and $3,000 for C-Cells. Of the 46,731 recruits available to fill 32,884 Rating slots, the baseline ASVAB composite set resulted in 19,702 A-Cells, 2,859 B-Cells, and 10,323 C-Cells for a total recruiting cost of $278,829,000. By comparison, the AO and CS augmented composite set resulted in 18,931 A-Cells, 2,703 B-Cells, and 11,250 C-Cells for a total recruiting cost of $271,734,000. Use of the augmented composite set with AO and CS resulted in a savings of $7,095,000 using the recruiting costs at the time (all 32,884 slots were filled in each case). Recruiting costs have grown over the years and the model outcomes could be updated if policy makers doubt the cost effectiveness of the two tests.

Concluding Remarks

The two Navy Rating assignment simulation applications are important tools that can be used during an ASVAB validation/standards study's cutscore setting process. As we know, changing an ASVAB standard for one Navy Rating is not done in a vacuum and we recognized there can be an impact on the fill and quality of recruits for the whole system of Ratings when other Rating's ASVAB standards are changed, and most importantly, when recruiting environments deteriorate. Typically stress in filling Ratings occurs only during a poor recruiting market (at the time of this writing, the recruiting environment seems to be slipping as Delayed Entry Program (DEP) time is now less than one year). As personnel research psychologists, we can help recruiting by continuing applied research efforts in the identification of selection tests that (a) improve the ASVAB's differential assignment capability, (b) reduce adverse impact, and (c) improve the predictive validity (or at the least, maintain it).

The next chapter expands on the much more complicated classification models developed for the "allocation" of personnel to military occupations – all of them striving to improve military classification effectiveness.

Chapter 18. References

Alderton, D. L., Wolfe, J. H., & Larson, G. E. (1997). The ECAT Battery, *Military Psychology, 9*, 5-37.

Carey, N. B. (1994). Computer predictors of mechanical job performance: Marine Corps findings. *Military Psychology, 6*, 1-30.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge, UK: Cambridge University Press.

Cattell, R. B. (1943). The measurement of adult intelligence. *Psychological Bulletin, 40,* 153-193.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (revised edition)*. Hillsdale, NJ: Erlbaum.

Crookenden, M., & Blanco, T. (2002). *SCORE test, comparing composite tests*. (Delivery Order EDS 02-002). EDS draft technical report prepared for Navy Personnel Research, Studies, and Technology, Millington, TN.

EDS Federal. (2001). *Alternative algorithms for job selection (*Contract Number: N66001-00-F-CR00/GS-35F-0015K Delivery Order: FISC NOR (0014); Sabre (GSA 01-002). EDS technical report prepared for Navy Personnel Research, Studies, and Technology, Millington, TN.

Held, J. D., & Carretta, T. R. (2013). *Evaluation of Tests of Processing Speed, Spatial Ability, and Working Memory for use in Military Occupational Classification* (NPRST-TR-14-1). Millington, TN: Navy Personnel Research, Studies and Technology

Held, J. D., Fedak, G. E., Crookenden, M. P., & Blanco, T. A. (2002). Test evaluation for augmenting the Armed Services Vocational Aptitude Battery. *Proceedings of the 44th Annual Conference of the International Military Testing Association,* pp. 291-297. Ottawa, Canada.

Held, J., Fedak, G., & Johns, C. (2004). *Armed Services Vocational Aptitude Battery (ASVAB) Standards: Aviation Mechanics Ratings* (NPRST Letter report Ser 3900, PERS-13/000011, 24 Feb 2004).

Held, J. D., & Wolfe, J. H. (1997). *Validities of Unit-weighted Composites of the Armed Services Vocational Aptitude Battery (ASVA B) and Enhanced Computer Administered Tests (ECAT). Military Psychology, 9,* 77-84.

Hogan, P., & Simonson, B. (2004a). *Selection and classification cost effectiveness model* (Contract Number: MOBIS N00639-02-F-A-097). The Lewin Group, Inc. software and user manual prepared for Navy Personnel Research, Studies, and Technology, Millington, TN.

Hogan, P., & Simonson, B. (2004b). *Selection and classification cost effectiveness model* (Contract Number: MOBIS N00639-02-F-A-097). The Lewin Group, Inc. technical report prepared for Navy Personnel Research, Studies, and Technology, Millington, TN.

Lawshe, C. H. (1987). Adverse impact: Is it a viable concept? *Professional Psychology, Research and Practice, 18,* 492-497.

Maier, M. H., & Sims, W. H. (1986). *The ASVAB score scales: 1980 and World War II* (CNR 116/July 1986). Alexandria, VA: Center for Naval Analyses.

Majeres, R. L. (1988). Serial comparison processes and sex differences in clerical speed. *Intelligence, 14,* 149–165.

Pommerich, M. (2013, October). *Preliminary evaluation of Coding Speed*. Presentation given to the Navy by Defense Manpower Data Center – Personnel Testing Division. Seaside, CA: Defense Manpower Data Center.

Sackett, P. R., Laczo, R. M., & Lippe, Z. P. (2003). Differential prediction and the use of multiple predictors: The omitted variables problem. *Journal of Applied Psychology, 88*, 1046-1056.

Scholarios, D., Johnson, C., & Zeidner, J. (1994). Selecting predictors for maximizing the classification efficiency of a battery. *Journal of Applied Psychology, 3,* 412-424.

Segall, D. O. (1997). The psychometric comparability of computer hardware. In W. A. Sands, B. K. Waters, & J. R. McBride, (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 219-226). Washington, DC: American Psychological Association.

Segall, D. O. (2010, October). *Table Processing Speed: A new processing speed test*. Presentation given to the Manpower Accession Policy Working Group. Seaside, CA: Defense Manpower Data Center.

Watson, S. E. (2004). The U.S. Navy's Rating Identification Engine (RIDE): Optimizing human resource allocation. *Proceedings of the 46th Annual Conference of the International Military Testing Association,* pp. 581-585. Brussels, Belgium.

Watson, S. E. (2010). Testing, validating, and applying an empirical model of human performance in a high-performance organization. In P. E. O'Conner & J. V. Cohn (Eds.) *Human performance enhancements in high risk environments: Insights, developments, and future directions from military research (technology, psychology, and health)* (pp. 316-336). Santa Barbara, CA: Greenwood.

Wise, L., Welsh, J., Grafton, F., Foley, P., Earles, J., Sawin, L., & Divgi, D. R. (1992). *Sensitivity and fairness of the Armed Services Vocational Aptitude Battery (ASVAB) technical composite (*DMDC Technical Report 92-002).

Zeidner, J., Johnson, C.,Vladimirsky, Y., & Weldon, S. (2004). *Comparison of alternative methods of measuring ASVAB test composite fairness* (ARI Study Note 2004-06). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Chapter 19.
Classification Effectiveness
Janet D. Held

Introduction

The last chapter described two simulation tools used by ASVAB validation/ standards researchers to assess the impact of changing an ASVAB standard for one Navy Rating (or an occupational group of Ratings) on the Navy's ability to fill all of their Ratings. The tools intentionally have simple understandable algorithms with only two inputs (ASVAB standards and recruits' ASVAB scores). The predictive validity of an ASVAB composite is not a parameter input, so it is up to the ASVAB validation/ standards researcher to fully understand all of the dynamics of a particular study and how to set an effective cutscore (Chapter 17). This chapter provides a discussion of some personnel allocation models developed for the military that are much more complicated, including the use of the validity coefficient, either explicitly or indirectly. We can consider these models important in selection and classification as they are intended to optimize classification gains, which we refer to as "classification effectiveness".

Military Service Work in Augmenting Utility Models

Chapter 3 provided a discussion on interpreting the validity coefficient in terms of utility (cost savings) to the organization from use of a valid selection or classification instrument. Chapter 18 provided a discussion of some simple personnel allocation models. Sands (1973) developed a more complicated model that captures many important military costs and outcomes. A Navy-developed model called the Cost of Attaining Personnel Requirements (CAPER) incorporated the Taylor-Russell (1939) tables discussed in Chapters 3 and 17. The CAPER model takes into account the magnitude of the validity coefficient and includes the following factors:

Quota = # of graduates required from training to report to the job

Base Rate = the observed success rate

Proportion Qualified = for graduates and failures separately at each score on the predictor

Cost_1 = cost of recruiting

Cost_2 = cost of selection

Cost_3 = cost of induction (processing)

Cost_4 = cost of training

Cost_5 = cost of erroneous acceptance (failures)

Cost_6 = cost of erroneous rejection (taking into account recruiting environments)

Although the validity coefficient is not explicitly listed in the Sands (1973) CAPER model, it is incorporated in the empirical cutscore outcomes (Base Rate, Proportion Qualified). Sands provided the necessary equations for trading off all of the costs, as well as the formulas for calculating the erroneous acceptance and rejection numbers (which can also be directly derived from the Taylor-Russell (1939) tables – discussed in Chapter 3 with a worksheet example provided in Appendix B).

Sands (1973) noted that although many assumptions underlying the model's accuracy might not be tenable, the model is a more complete representation of what policy makers must evaluate to judge their selection systems and states:

> "In the tradition of classical test theory, the correlation model focuses upon the accuracy of measurement. In contrast, the CAPER model is decision oriented and recognizes the necessity of taking into account the utility or cost of various decision-outcome combinations" (p. 226).

As with some utility models, the CAPER model involves a single occupation and thus does not take into account that the military must fill many occupations with able recruits. Nevertheless, the CAPER model addresses the need to efficiently evaluate which in an array of cutscores applied to predictors produces the least cost to the personnel system – including cost factors through training. Other military classification systems such as the Navy's former Classification and Assignment within PRIDE (CLASP) (Kroeker & Folchi, 1984; Kroeker & Rafacz, 1984) take the operational cutscore as fixed and predetermined through the ASVAB validity/standards studies.

The Air Force conducted research using the Taylor-Russell (1939) tables to extend the model to more than one job (Alley, Darby, & Cheng, 1996). However, the research was constrained to a single selection ratio (reflected in the cutscore) limiting the method's use. Cheng and Darby (1997) also studied the classification efficiency-problem by extending the Brogden (1959) table of allocation benefits to more than one job. The assumption, however, was that the jobs were equally correlated (based upon the jobs' predicted performance scores). Alf and Abrahams (1996) for the Navy extended the Brogden allocation table to deal with up to 1,000 jobs and with a wide range of "rejection" rates (0 to 90%). However, as with all of the approaches just discussed, the cutscore setting process did not consider (a) the benefit or impact of a particular job to the organization (clerical vs. cyber network troubleshooting) or (b) the cost of training (e.g., one year for a Nuclear Field occupation vs. three months for an Administrative occupation).

For the Army, Schmitz and Holz (1987) reviewed the person-job matching problem from an operations research and computer science perspective but recognized that the optimum person-job matching algorithm involves an interdisciplinary team that includes personnel psychologists. Therefore, Schmitz and Holz incorporated into their model aspects of selection, differential classification, and assignment, making note of the simplicity in most job assignment algorithms as they evaluate each job separately and then aggregate gains to the organization (productivity) across jobs (p. 441) in a batch strategy. In reality, person-job fit and assignments are made sequentially, and it is

not assured that candidates further out will be sufficiently qualified for critical jobs. Conversely, a marginally qualified applicant may be accepted for the Nuclear Field on a particular day not knowing that one week later a more qualified applicant could have filled that final year's allocated contract.

Schmitz and Holz (1987) noted the increasing need at the time for more predictors that measure the important multidimensional criterion space of each job and also the recognition that there are differences in selection and classification:

> "Selection focuses on the differences among individuals, generally using a single scale of value or utility. Applicants are classified into two categories: those satisfactory for employment and those not….
>
> Differential classification deals with differences within an individual with respect to various skills. A particular individual may have a high aptitude for mathematics but poor writing skills. Another may have considerable talent for electronics jobs but poor communication ability. Classification requires the use of two or more different performance predictors" (p. 440).

We note here that predicting both training and on-the-job performance are important goals for the military because there are costs associated with personnel failures at either point. And, even though the ASVAB was developed to predict training performance, it does predict job performance to some extent because the ASVAB tests were developed from the linkage of training Knowledge, Skills, and Abilities (KSAs) requirements that are fundamentally derived from job analysis. The extent of ASVAB's prediction of job performance depends upon what aspects of job performance are measured. For example, the ASVAB predicts measures of job knowledge fairly well, but not organizational citizenship or discipline. The next section provides a discussion about the "goals" of a prediction system, and what critical performance constructs one needs to consider as the predicted criterion variable.

Goals for Improving Classification Decisions

Improving military enlisted classification decisions in general has been a topic of interest for a long time and has been viewed in the context of selection as well as classification. One consideration for quantifying what is an "improvement" would be to first identify what are the improvement goals. Rosse, Campbell, and Peterson. (2001, p. 456) highlight Wise's (1994) description of numerous potential military goals for improving selection and classification decisions, not all of which can be addressed by adding new cognitive tests to the ASVAB (non-cognitive measures are now being considered by the military services). Among Wise's stated goals, three are directly pertinent to setting of ASVAB classification standards for Navy classification. The first goal is to maximize the percentage of training seats filled with qualified applicants, which improves with lowering the average ASVAB test intercorrelation – that is, adding tests that uniquely measure person and job attributes for clusters of jobs that are different in job dimensions than others (just as we saw occurred in the last chapter from use of the Assembling Objects and Coding Speed tests).

The second goal for improving classification that applies directly to Navy ASVAB validation/standards studies is to improve the rates of training success. Training failure is a costly loss for the Navy and as we have seen in Chapters 3 and 17, we can mitigate those costs by improving ASVAB classification composite validities and setting effective cutscores. The third goal pertinent to the Navy (and the other Services) is to improve the social benefits, which means lowering adverse impact occurring from overuse of highly academic tests that are not the *only* credible predictors of training and occupational success. We note here DoD's current efforts in considering measures of personality, working memory, and non-verbal reasoning, and other constructs as additions or adjunct to the ASVAB (cited in Chapter 1 of the Introductory Manual.

Rosse et al. (2001) recognized Wise's (1994) position that it is impossible to simultaneously optimize all of the many possible selection and classification goals. It becomes a complicated matter for an organization to choose which goals to emphasize because they all are worthy. Table 19-1 lists the array of classification goals from Wise.

Table 19-1
List of Selection and Classification Goals (Wise, 1994)

1: Training seat fill	**7:** Total career performance
2: Training success	**8:** Total MOS performance
3: Attrition reduction	**9:** Performance utility
4: Job proficiency	**10:** Unit performance/readiness
5: Job performance	**11:** Adverse impact reduction
6: Months of service qualified	**12:** Preference accommodation

As we see from Table 19-1, there are several potential classification goals (and more not listed, such as retention) addressed by Wise (1994). In the military context, a predictor screen that would predict the goal of attrition reduction may not correlate with the ASVAB so, in effect; we would be layering a cutscore screening system that systematically reduces the military eligible youth population. We recognize that attaining selection and classification goals further out in time in an upfront selection and classification system would also be influenced by organizational factors such as pay and benefits, promotion potential, organizational climate, quality of the leadership, and possibly the propensity of youth to enlist (maybe more so in an economic downturn when jobs are scarce, but also through periods of high patriotism). These organizational factors can be "moderators" of the ASVAB's relation to performance, especially the further out goals. The risk of incorporating predictors of further out goals is that they influence performance predictions in some recruiting periods and not others.

We refer the reader to Laurence and Hoffman (1993) for an evaluation of the Services' classification systems and to Schmitz and Holz (1987) and move on to discuss differential assignment capability and classification efficiency.

Differential Assignment and Classification Efficiency

Of all the Services, the Army has been the most involved during past years in conducting research that shows improvements in classification considering the optimal person-job match. Classification improvement or classification efficiency (CE) considers both increased predictive accuracy from (a) full least squares (FLS) predictor regression equations and (b) the unobstructed ability to differentially classify individuals to their best-fit occupation (which, in reality, we do not have in the operational day-to-day classification of enlisted recruits).

Horst (e.g., 1954, 1955, 1956) and Brogden (1946, 1951, 1955, 1959) were early pioneers of classification effectiveness and CE, and the Army adopted their methods with further application. Johnson and Zeidner (1991a; 1991b, 1995) developed the principles of "Differential Assignment Theory" that considered the tenets of Brogden' s (1959) measure of CE capturing both predictive validity and the intercorrelation of the ordinary least squares (OLS) equation estimates of performance (job or training). The formulas are presented succinctly by Statman, Gribben, Naughton, and McCloy (1998, p. 7) as follows:

$$M_{PP} = R\,(1\text{-}r)^{1/2}\,Z_m \text{ where,}$$

M_{pp} = the mean predicted performance standard score of a group of applicants assigned to m jobs,

R = the average predictive validity of ordinary least squares (OLS) estimates for all jobs,

r = the average intercorrelation of the OLS estimates, and

Z_m = the mean criterion standard score of the group after assignment to the m jobs with equal vacancies (called quotas).

Statman et al. (1998, p.7) point out clearly that R is the predictive validity function that is positively related to CE but that to maximize the CE index of $M_{PP,}$ one needs to lower the intercorrelations of the prediction equations across occupations. Lowering the average intercorrelation of the ASVAB tests would enable this goal as did the AO and CS tests discussed in the previous chapter.

The Air Force sponsored interesting work conducted by Statman et al. (1998) that involved transporting optimal classification methods to training evaluation and fit. The point was that, as with occupations, not all individuals would be expected to thrive in all training formats. The Statman et al. work is of high interest as the Navy and other Services deal with iterations of training formats/platforms in a resource constrained fiscal environment.

We direct the reader to Scholarios, Johnson, and Zeidner (1994) for a capsulated presentation of how differential assignment theory (DAT) should be considered when forming ASVAB classification composites and the implications for augmenting the ASVAB to make it more classification effective. We also note that Schmidt, Hunter, and Dunn (1987) showed cost savings for the Navy by adding a perceptual accuracy and a psychomotor test to the ASVAB. With regard to the former ASVAB Coding Speed (CS) test (now a special classification test for Navy), Scholarios et al. (1994) studied optimal test battery composition for the Army (from Project A measures). It is interesting that in a full-least squared regression equation, the CS test entered in first place under the specified objective of optimizing differential assignment capability (for the instance of 18 jobs – the most Army occupations included in the study). It is also interesting that over time, CS test scores have been shown to relate to earnings level for moderately complex civilian occupations, like many in the military, and that most likely; part of the reason is due to measurement of a degree of intrinsic motivation (Segal, 2012).

Under the assumption of job relevance and uniqueness, the Enhanced Computer-Administered Test (ECAT) psychomotor test, Two-Hand Tracking, showed incremental validity to the ASVAB in the late 1980/early 1990 time frame for certain Army tank jobs (Abrahams et al., 1993; Wolfe, 1997). Others have found that psychomotor tests measure a relevant unique construct but only trivial incremental validity to the ASVAB - most likely because of the mismatched occupation (mechanical) (Mayberry & Divgi, 1992). Psychomotor tests, although good candidates for military classification, have traditionally required computer hardware peripherals that would not be cost effective to maintain in large-scale testing programs. Getting past the need for computer peripherals to measure psychometric ability seems to be a fruitful military classification research project (email communication from Dr. Phillip Ackerman April 6, 2013).

In closing, to further make the case for ASVAB tests that add differential assignment capability and may by their nature increase classification effectiveness and CE (as measured by mean predicted performance (MPP), Abrahams et al. (1994) showed that some of the tests in the ECAT battery added utility to the ASVAB (focus on Brogden's [1959] measure of CE that incorporated the differential assignment function). Also, Alley and Teachout (1995) applied a linear programming algorithm with optimally-weighted ASVAB/experience variables for a constrained number of Air Force job assignments comparing Least Squares Estimates (LSEs) with the then current operational assignment method, but under a random assignment strategy. The improvement in the average expected performance gains over the baseline was one third of a standard deviation, which translated to 14 months of technical experience.

Both the Army and Air Force have used measures of job performance as the criterion for their research. The Air Force performance criterion was developed as in-depth hands-on tasks/procedures tailored to each of the training specialties included in the study. Although the research did not focus on CE as measured by MPP, the results suggest, along with Army research, that there is merit for the Services to consider their full ASVAB regression equation development as part of the DAT principles. Alley (1994) discussed advancements at an earlier time in classification from the Air Force perspective.

Concluding Remarks

In this chapter, we discussed various methods (algorithms) applied to sometimes different variables for the purpose of optimizing the classification (and assignment) of people to jobs. The chapter was intended to give a broader context to ASVAB validation/standards studies and the final threading together of how selection and classification tests in the military personnel setting are evaluated and valued (utility). The next chapter summarizes the key points in each of the technical manual chapters.

Chapter 19. References

Abrahams, N. M., Kieckhaefer, W. F., Cole, D. R., Alf, E. F., Pass, J. J., & Walton-Paxton, E. (1994). *Classification utility of test composites from the ASVAB, CAT-ASVAB, and ECAT batteries* (Contract No N66001-90-D-9502 Delivery Order 7J16). San Diego, CA: Navy Personnel Research and Development Center.

Abrahams, N. M., Pass, J. J., Kusulas, J. W., Cole, D. R., Kieckhaefer, W. F., Humphreys, L. G., & Alf, E. F. (1993). *Incremental validity of experimental computerized tests for predicting training criteria in military technical schools* (Contract No N66001-90-D-9502 Delivery Order 7J13). San Diego, CA: Navy Personnel Research and Development Center.

Alf, E. F., & Abrahams, N. M. (1996). The impact of number of jobs and selection ratio on classification efficiency: An extended table of Brogden's allocation average. *Educational and Psychological Measurement, 56,* 951-956.

Alley, W. E. (1994). Recent advances in classification theory and practice. In M. G. Rumsey, C. B. Walker, & J. H. Harris (Eds.), *Personnel selection and classification* (pp. 431-442). Hillsdale, NJ: Erlbaum.

Alley, W. E., & Teachout, M. S.(1995). *Differential assignment potential in the ASVAB: A simulation of job performance gains* (AFHRL-TP-1995-0006). Brooks Air Force Base, TX: Human Resources Directorate, Armstrong Laboratory.

Alley, W. E., Darby, M. M., & Cheng, C. (1996, October). *The practical benefits of personnel testing: An extension of the Taylor-Russell tables to multiple job categories*. The 38th Annual Conference of the Military Testing Association, San Antonio, TX.

Brogden, H. E. (1946). An approach to the problem of differential predictions. *Psychometrika, 11*, 139-196.

Brogden, H. E. (1951). Increased efficiency of selection resulting from replacement of a single predictor with several differential predictors. *Educational and Psychological Measurement, 11*, 173-196.

Brogden, H. E. (1955). Least squares estimates and optimal classification. P*sychometrika, 20, 249-52*.

Brogden, H. E. (1959). Efficiency of classification as a function of number of jobs, percent rejected, and the validity and intercorrelation of job performance estimates. *Educational and Psychological Measurement, 19*, 181-190.

Cheng, C., & Darby, M. (1997). *Efficiency of classification: Revision of the Brogden Table* (AL/HR-TP-1997-0013). Brooks AFB, TX: Manpower and Personnel Division, Armstrong Laboratory, Human Resources Directorate.

Horst, P. (1954). A technique for the development of a differential predictor battery. *Psychological Monographs, 68,* 1-31.

Horst, P. (1955). A technique for the development of a multiple absolute prediction battery. P*sychological Monographs, 69,* 1-22.

Horst, P. (1956). Multiple classification by the method of least squares. *Journal of Clinical Psychology, 12,* 3-16.

Johnson, C. D., & Zeidner, J. (1991a). *The economic benefits of predicting job performance Vol II: Classification efficiency*. NY: Praeger.

Johnson, C. D., & Zeidner, J. (1991b). *The economic benefits of predicting job performance Vol III: Estimating the gains of alternative policies*. NY: Praeger.

Johnson, C. D., & Zeidner, J. (1995). *Differential assignment theory sourcebook* (ARI Research Note 95-43). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Kroeker, L. K., & Folchi, J. S. (1984). *Classification and assignment within PRIDE (CLASP) system: Development and evaluation of an attrition component* (NPRDC-TR-84-40). San Diego, CA: Navy Personnel Research and Development Center.

Kroeker, L. K., & Rafacz, B. A. (1984). *Classification and assignment within PRIDE (CLASP): A recruit assignment model* (NPRDC-TR-84-9). San Diego, CA: Navy Personnel Research and Development Center.

Laurence, J. H., & Hoffman, G. R. (1993). *A description and evaluation of selection and classification models* (HumRRO Technical Report 979). Alexandria, VA: Human Resources Research Organization.

Mayberry, P. W., & Divgi, D. R. (1992). *Uniqueness of a psychomotor construct to the ASVAB* (CRM 91-224/January 1992). Alexandria, VA: Center for Naval Analyses.

Rosse, R. L., Campbell, J. P., & Peterson, N. G. (2001). Personnel classification and differential job assignments: Estimating classification gains. In J. P. Campbell & D. Knapp (Eds.), *Exploring the limits in personnel selection and classification (pp. 453-506)*. Mahwah, NJ: Earlbaum.

Sands, W. A. (1973). A method for evaluating alternative recruiting selection strategies: The CAPER model. *Journal of Applied Psychology, 57,* 222-227.

Schmidt, F. L., Hunter, J., & Dunn, W. (1987). *Potential utility increases from adding tests to the Armed Services Vocational Aptitude Battery (ASVAB)* (NPRDC-TN-95-5). San Diego, CA: Navy Personnel Research and Development Center.

Schmitz, E. Z., & Holz, B.W. (1987). Technologies for person-job matching. In J. Zeidner (Ed.), *Human productivity enhancement* (Vol 2, pp.439-472). NY: Praeger.

Scholarios, D., Johnson, C., & Zeidner, J. (1994). Selecting predictors for maximizing the classification efficiency of a battery. *Journal of Applied Psychology, 3,* 412-424.

Segal, C. (2012). Working when no one is watching: Motivation, test scores, and economic success. *Management Science, 58,* 1438-1457.

Statman, M. A. L., Gribben, M. A., Naughton, J. A., & McCloy, R. A. (1998). *The development of a new research paradigm for studying aptitude-treatment interactions* (HumRRO FR-WATSD-97-25). Alexandria, VA: Human Resources Research Organization.

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology, 23,* 565-578.

Wise, L. (1994). Goals of the selection and classification decision. In M. G. Rumsey, C. B. Walker, & J. H. Harris (Eds.), *Personnel selection and classification (pp. 351-361).* Hillsdale, NJ: Erlbaum.

Wolfe, J. H. (1997). Incremental validity of ECAT battery factors. *Military Psychology, 9,* 49-76.

Chapter 20.
Summary of Key Chapter Points and Future Concerns
Janet D. Held

The intent of the Introductory and Technical ASVAB Validation/Standards manuals is to provide comprehensive guidelines for sponsors, policy makers, and researchers in conducting ASVAB validation/standards studies. It is recognized that not all statistical methods associated with criterion related test validation have been addressed in the manuals, but that what is provided serves as a sound basis for further exploration. We also note that it is not the ASVAB per se that we validate, but our use of the scores. Unfortunately, because the ASVAB is not a perfect predictor of military performance, there will always be a tradeoff in how many recruits qualify for the Services' occupations and the success rates of those recruits in training, all in the context of changing military recruiting environments and fluxuations in training resources. This chapter summarizes some of the key points in the chapters of each manual in hopes of presenting a logically flowing coherent bigger picture of why and how we set ASVAB standards.

The Introductory Manual

Chapter 1: Introduction. This chapter established that the purpose of the Introductory Manual is intended to provide general guidance to policy and researchers on conducting ASVAB validation/standards studies. The main Navy stakeholders are the sponsors of the program, the Navy's Selection and Classification Office, Navy Recruiting Command, Navy Training, and Enlisted Community officials who manage their Rating communities and who develop policy. Ultimately the Fleet is also a stakeholder as the integrity of all three of these entities benefit from effective ASVAB standards. The Technical Manual (volume 2) is introduced and its purpose is described.

Chapter 2: An Overview of the ASVAB. This chapter provided a brief history of the ASVAB and a description of the aptitudes, abilities, and knowledge constructs that the ASVAB measures. It also described a large-scale joint-service test development and validation effort that was intended to augment the ASVAB with tests that are more representative of fluid intelligence (i.e., the ability to reason abstractly and solve problems) rather than crystallized intelligence (i.e., the ability to apply accumulated knowledge, skills, and abilities). A list of criteria were proposed for use in evaluating candidate tests for the ASVAB noting that recent revisions to this list have been made and approved by the Manpower Accession Policy Working Group (MAPWG) that oversees ASVAB development.

Chapter 3: Mapping the predictors and Criteria. This chapter was a discussion of a framework for understanding the predictor and criterion relations as overt measures captured by measurement instruments, but also of their underlying constructs. The chapter distinguished the military's operationally-oriented setting of ASVAB standards from theory-based research.

Chapter 4: Issues in Predicting Job Performance. This chapter highlighted issues involved in predicting job performance, including the sparse literature on the criterion compared to that dedicated to predictor development. A discussion of the well-known but seldom-addressed "Criterion Problem" and the difficult decisions about whether to consider single or multiple criteria was provided.

Chapter 5: Navy Training and Best Practice Performance Measurement. This chapter focused on the criterion as training performance and described the evolution of Navy training from instructor-led group-paced classroom courses with hands-on laboratory demonstrations to self-paced computer-based training. Guidelines are provided for best practice performance measurement in training, in particular for simulation-based training. Also provided are lessons learned from the literature on successful and unsuccessful computer-based training, which has guided the Navy's evolution to a blended training solution.

Chapter 6: The Navy's ASVAB Validation/Standards Process. The process and steps for conducting Navy ASVAB validation/standards studies were laid out. Although the process generalizes for each Navy Rating, each study is tailored to the particular issues that either directly or indirectly relates to the effectiveness of a Rating's ASVAB standard. The overall goal of any ASVAB validation/standards study is to provide standards that minimize academically-related failure and setback rates (that result in high costs to both the Navy and individual) while at the same time addressing the Navy's need to fill all jobs with qualified Sailors.

Chapter 7: Applications of Synthetic Validity. An overview was provided of an indirect approach to estimating test validity when sample sizes are too small for a robust analysis or, in the case of the Navy, a new occupation (Rating) is stood up and researchers do not yet have performance data. Synthetic validity is in the class of validity generalization methods. The Navy has its own approach to suboptimal validation situations; however, synthetic validity is of interest and various aspects of it may be considered as part of the ASVAB validation/researchers tool box in the future if performance data for one reason or another become generally unavailable.

Appendices A, B, and C of the Introductory Manual contain examples of Navy ASVAB validation/standards studies to demonstrate that one size does not fit all. That is, the three studies (Navy SEALs, Nuclear Field, and Mineman Ratings) are provided to (a) demonstrate the dynamic issues that can differ for each Navy rating, (b) show how the study methods are tailored to address them, and (c) demonstrate the wide range of ASVAB composite predictive validity coefficient magnitudes.

The Technical Manual

Chapter 1: Introduction. This chapter was a discussion about the need to develop general and specific technical guidance for the Navy and the other military services in conducting ASVAB validation/standards studies. Each of the Services plays a major role in improving the selection and classification outcomes of their enlisted members through the Manpower Accession Policy Working Group (MAPWG). The MAPWG deals with the ASVAB, with each Service taking major roles in different related areas. For

example, the Army is the lead in demonstrating the value of non-cognitive measures and the Air Force in the development of a Cyber Test. The Navy has the lead in developing ASVAB validation/standards guidance and the Defense Manpower Data Center, Personnel Testing Division (DMDC-PTD) as ASVAB Executive Agent, develops and maintains the battery and hosts/monitors all new candidate ASVAB tests.

Chapter 2: Predictor-Criterion Relations: A Brief Statistical Overview. A brief statistical overview was presented of correlation and regression analysis. Regression analysis is used to identify a linear trend in the data, and the correlation coefficient is used to quantify the fit of the data to the trend. We discussed the importance of creating a scatter plot to examine the data for outliers and nonlinearity and ways to test for statistical significance with one or more predictors. We observed that the precision of estimates of r, b, and y depend on the size of both the correlation coefficient and the sample size in addition to other factors that affect correlation and regression analysis. We also reviewed classical test theory, measurement error, confidence intervals, restriction in range of test scores used in selection and the need to correct for range restriction due to the selection/classification standard. We also reviewed ways to estimate associated errors in the resulting corrected correlation (validity coefficient) recognizing that sample size is an important factor.

Chapter 3: Interpreting the Correlation (Validity) Coefficient. This chapter was about interpreting the validity coefficient and the utility of a valid selection instrument. For the Navy, the ASVAB validity coefficient is developed using the continuous final school grade variable as the criterion whereas training completion (pass/fail) status is used for cutscore analysis. Interpreting the benefits of a valid selection instrument typically has been shown over random assignment, but in the military case involving the ASVAB, the improvements are shown from optimizing the combination of ASVAB tests used in an occupational classification composite. In this case, utility analysis can address benefits as the gain in average expected improvements in performance from replacing a suboptimal composite. Another way to interpret the validity coefficient is in terms of classification decision accuracy where, all other things being equal, the magnitude of the validity coefficient generally will improve the likelihood of making correct decisions (selecting individuals who succeed and not selecting those who would fail).

Chapter 4: Measurement Error and Reliability Estimators. This chapter reviewed measurement error and the derivation of the formulas that lead to a correction for less than perfect reliability. We discussed the types of reliability estimation methods and, to broaden the context, two common ways of measuring job performance (interrater and intrarater). Examples from the literature highlighted the appropriate conceptions and applications of a correction for measurement error. ASVAB reliabilities were cited for paper-and-pencil forms with a DMDC website link for information about the CAT-ASVAB method of reliability estimates. We also provided reliabilities from the military Job Performance Measurement (JPM) project and some from meta-analysis of non-military developed job performance criteria. We acknowledge that in setting ASVAB standards, the Navy does not correct either the ASVAB or the training performance criterion for unreliability because of the operational focus on enlistment qualification/standards setting.

Chapter 5: Correcting for Restriction of Test Score Range. Derivations were provided for correcting the validity coefficient for explicit and incidental restriction in range. Restriction in range of test scores occurs when, from the application of a selection/classification standard, the result is to curtail variance and therefore the correlation (validity) coefficient. We show an example of the biasing effect of range restriction taken from an early Army Air Force study. The Navy, in validating the operational and candidate replacement ASVAB composites for a specific Navy Rating, applies the multivariate correction for range restriction, also described in the chapter, because all ASVAB test scores are available for the normative and applicant populations.

Chapter 6: Joint Corrections for Measurement Error and Range Restriction. This chapter reviewed the basic underpinnings of the correction for measurement error and range restriction as review and background for the complicated joint correction. The order of the joint correction is an important consideration and the availability of reliability estimates has typically been the determinate. We stressed that the joint corrections are most appropriately applied in theory-based research when the objective is to estimate the underlying relation between variable constructs.

Chapter 7: More on Joint Corrections. This chapter brought a slightly different perspective to the joint corrections and showed that in the case of direct/explicit selection, the unrestricted reliability of the explicit selection variable is downwardly-biased. The downward bias is the result of a negative correlation between true and error scores that, from classical test theory, is assumed to be zero. A remedy is developed that involves (a) retest of the explicit selection test in the selected sample and (b) formula modifications. The chapter stresses that the criterion and predictor reliability must be brought to the same level for the joint correction and that the preferable sequence is to correct first for restriction in range, and then for reliability (assuming unrestricted reliabilities are available).

Chapter 8: Standard Errors of the Corrected Correlation. This chapter provided a brief review of the two major approaches reported in the literature for estimating standard errors of corrected correlations due to range restriction: (a) asymptotic sampling variance formulas and (b) bootstrapping. The formula approach is based on asymptotic sampling variance theory and therefore the standard error can be used for computing confidence intervals or testing for significance. The bootstrap is a non-parametric approach not tied to distribution assumption.

Chapter 9: A Monte Carlo/Bootstrap Study of Range Corrected Validity Accuracy. Some key findings were reported from an ASVAB Monte Carlo simulation study with the bootstrap method applied to the Monte Carlo generated samples. The objective was to study the accuracy of the multivariate range correction procedure under varying study conditions – selection ratio, predictor/criterion skew, and sample size. Two ASVAB tests served as surrogate training criteria. Over and underestimates of the population validities were observed with degree of bias related to extremes in the conditions. Most notably for ASVAB validation/standards studies, the larger skew values imposed on the criterion resulted in large validity bias. However there was small bias, if any, when the skew was applied progressively in magnitude to the ASVAB predictors, reinforcing the importance of the psychometric integrity of the criterion variable.

Chapter 10: Assumption Violation Effects on Range Corrected Accuracy. A discussion was provided on the nature of violations in the assumption for performing the correction for range restriction. Assumption violations can lead to largely biased estimated population validity coefficients, or they offset each other and yield accurate results. Even in the best of circumstances, however, we can never accurately extrapolate a sample's characteristics to the unrestricted population to assess assumption violations. There is, however, more certainty about the situation when the selection ratio is large (lenient selection) than when it is small – most of the Services' selection ratios for their occupations are not extremely stringent – the exception is for jobs with high aptitude requirements such as the Navy Nuclear Field and crypto linguist specialties.

Chapter 11. The Potential for a Negative Range Corrected Validity. This chapter described the conditions under which a validity coefficient's sign changes when corrected for range restriction. The negative-to-positive sign change is a usual expectation with very stringent selection ratios. On the other hand, the positive-to-negative sign change is unusual and is a function of a small data set, severe selection stringency, and use of multivariate range restriction formulas (that allows regression weights to wildly offset each other). The positive-to-negative sign change should be viewed as an unrealistic outcome and suggests that in ASVAB validation/standards studies, we should evaluate our sample as best we can, including an examination of the full least squares regression weights that are applied in the multivariate correction.

Chapter 12. Partial Correlation, Hierarchical and Logistic Regression, and Power. Two regression methods were reviewed – hierarchical and logistic –often used in predictor/criterion research and test validation studies. ASVAB validation/standards studies utilize these methods selectively dependent upon the objectives and the nature of the available performance variables. Statistical power was explored as a complicated requirement for planning sample size when the ASVAB data are restricted in range due to an ASVAB standard.

Chapter 13. Weighting Variables: The Tradeoff between Validity and Adverse Impact. A general discussion was provided of methods for establishing weights for the independent variables in a prediction equation, such as least-squares regression weights versus relative weights. The literature shows mixed results about the conditions under which each method should be used. There are distinct advantages that "relative importance" indices have over regression coefficients in terms of communicating to decision makers. There are also advantages to just using integer weights when the sample size is small as they generalize better to new samples. The Navy and other Services except the Army use integer (unit) weights to form their ASVAB composites.

Chapter 14. More on Weights: Forming a Composite of Multiple Performance Criteria. The discussion was continued about methods for applying weights with focus on the "complete" criterion. Guidance was given with regard to distinguishing between desired weighting of the components of a composite ("nominal weights") and the weights that actually are in play because of variances of and covariances among the components ("effective weights"). Equations were presented that allow one to calculate appropriate empirical weights that consider the variances and covariances of the components to yield effective weights that match the desired nominal weights.

Chapter 15. Multiple Hurdles and the Correction for Range Restriction. A discussion was provided of multiple hurdle selection systems and a two-step correction for range restriction procedure to estimate the unrestricted population predictor/criterion correlation matrix. The application for ASVAB validation studies is when (a) the ASVAB is used to predict performance in advanced training (intermediate training needs to be taken into account) and (b) a researcher decides job performance is the ultimate criterion (training performance needs to be taken into account). Not accounting for a hurdle will result in a downward bias in the ASVAB's estimated population validity coefficient. The 2-step procedure assumes the missing performance data can be filled in by some legitimate procedure.

Chapter 16. Multiple Hurdles as a Missing Data Problem. This chapter drew some parallels to the multiple hurdle range correction issue discussed in the previous chapter under the umbrella of "the missing data problem". A major point stressed is that the missingness of data, either in a one or two-stage hurdle situation, is missing at random (MAR). That is, the missingness must be related only to the selection variable (e.g., the ASVAB) and not to the criterion variable after controlling for the ASVAB; otherwise, we have missing not at random (MNAR) and our model will be misspecified. Maximum Likelihood and the state-of-the-art Multiple Imputation procedures (that provide standard errors) are briefly discussed, but they too, as with the Lawley 2-step correction, must adhere to assumptions.

Chapter 17. Setting ASVAB Cutscores. A brief background was provided from the literature on methods for setting cutscores and also the impact of measurement error on selection decisions. Multiple cutscores were discussed in contrast to compensatory composite models and multiple-hurdle models. Also discussed was the use of empirically-based expectancy tables compared to theory-based tables, the latter are required for the Navy when an ASVAB cutscore is to be lowered (or waived), or the sample size is too small but the validity estimate seems reasonable. Waiver policy guidelines are provided for the Navy that takes into account various factors such as criticality of the Rating, difficulty in filling the Rating, ASVAB validity magnitude, and observed training failure rates.

Chapter 18. Assessing ASVAB Standards Adequacy through Simulation. This chapter was a discussion of two Navy simulation-based software applications that allow for the assessment of impact of recruit fill across Navy Ratings when the ASVAB standard is changed for one or more Ratings (or an occupational group of Ratings). Both applications were applied in a study of the differential assignment capability of the newest ASVAB test, Assembling Objects (AO) and the former ASVAB test, Coding Speed (CS). Both tests have been shown to increment the ASVAB in predicting performance in mechanical types of occupations (for AO) and clerical, but also other types of Ratings including Navy SEALs (for CS). The results for both applications show that these tests, that also reduce adverse impact, improve the fill across Navy Ratings when formed into ASVAB composites that have shown optimal validity.

Chapter 19. Classification Effectiveness. This chapter expanded upon the simple classification algorithms discussed in the previous chapter to the more complex and complete recruit classification/allocation models researched by the Services'. These models attempt to improve classification effectiveness. A discussion is provided about the various selection and classification goals that are so varied that they complicate the matter of deciding what any revised ASVAB should predict. The Navy has three main goals for the ASVAB to accomplish through their Ratings' ASVAB standards. The first goal is to maximize the percentage of training seats filled with qualified applicants. The second goal is for the ASVAB to be highly predictive of training success – as failing training is a costly loss for the Navy (and the individuals). The third goal for the ASVAB is to improve the social benefits, which means lowering adverse impact, or score barriers, that occur from overuse of highly academic tests that are not the only credible predictors of training success.

Appendix A: SPSS syntax file that executes the Pearson-Lawley correction for range restriction and the associated files and instructions. Appendix B: A generated Taylor-Russell .10 base rate table. Appendix C: Worksheet to demonstrate how to use the Taylor-Russell tables to determine the four classification decision errors.

Concerns for the Future

During the writing of the Introductory and Technical manuals, the military was experiencing an extended positive military recruiting environment (juxtaposed to a poor U.S. jobs market). Military recruiting environments are cyclical and a downturn is expected and we are already seeing downturn signs, and it is occurring at a time when the U.S. government is experiencing large budget deficits. Military funding cuts will impact training budgets, so we in the ASVAB validation/standards community will not be assured of the stellar criterion measures we currently receive from the training commands. If training resources are severely cut, our training performance criterion variables may become less useful.

Despite resource constraints the DoD ASVAB program will remain high in integrity; however, we caution against researchers just settling for performance criteria of convenience to validate the ASVAB. We must be diligent in investigating the construct relevance and other criterion properties addressed in these two manuals (Introductory and Technical). If for some reason training performance measures become unavailable, some could rationalize that the criterion of relevance for setting ASVAB standards is, say promotion status – fully documented in electronic databases and easy to obtain with the appropriate DON Chief Information Officer and IRB approvals. To counteract any counterproductive developments that may occur for the function of ASVAB standards setting, it will become imperative that the Services leverage the two manuals to educate policy makers on the importance and utility of our standards setting practices. At least for the Navy, there exists a Selection and Classification Office (OPNAV132G) that formalized the ASVAB Validation/Standards program as an operational requirement. The other Services may find it beneficial to do the same. Also key is the sustainment of the INTERSERVICE Aptitude/Ability Standards Panel (see the charter in Appendix D of the Introductory Manual, NPRST-TR-15-1).

Appendix A
Multivariate Correction for Range Restriction Procedures and SPSS files

SPSS SYNTAX as examples for setting up RANGE RESTRICTION CORRECTION matrices

```
*COMPUTE ASVAB COMPOSITES TO CORRECT FOR RANGE RESTRICTION
COMPUTE vear= ve+ar.
COMPUTE vemk= ve=mk.
COMPUTE ar2mkgs= ar+(2*mk)+gs .
EXECUTE .


*GENERATE THE SAMPLE MATRIX
CORRELATIONS
/VARIABLES=gs ar wk pc mk ei as mc ve ao vear vemk ar2mkgs fsg
/MATRIX OUT (*).


*REGRESSION PROGRAM APPLIED TO RANGE CORRECTED MATRIX
REGRESSION
/MATRIX=IN(*)
/MISSING LISTWISE
/STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT fsg
* The Lawley corrected matrix yields range corrected validities for unit weighted ASVAB test composites
*where test scores are standardized in the PAY97 population to have a mean of 50 and
*standard deviation of 10. The composites are the ones you specify in the Cormat.spx file that sets up
*The composites are those you specify in the Cormat.spx file that sets up the samp_mat.sav file.
*The researcher may want to run regression on the corr_mat file as well as directly on the samp_mat (or the
*sample's raw data) for a number of reasons. For example, R-Square is only meaningful in the corr_mat
*recognizing that accuracy is influenced by a number of factors addressed in the Technical Manual.

/METHOD=ENTER gs ar wk pc mk ei as mc ve ao.
```

SPSS SYNTAX as examples for conducting regression on the RANGE RESTRICTION CORRECTION matrix

```
*REGRESSION PROGRAM APPLIED TO RANGE CORRECTED MATRIX
REGRESSION
/MATRIX=IN(*)
/MISSING LISTWISE
/STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT fsg
*fsg is final school grade.
* The Lawley corrected matrix yields range corrected validities for unit weighted ASVAB test composites
*where test scores are standardized in the PAY97 population to have a mean of 50 and
*standard deviation of 10. The composites are the ones you specify in the Cormat.spx file.
*The researcher may want to run regression on the corr_mat file as well as directly on the samp_mat (or the
*sample's raw data) for research reasons.

/METHOD=ENTER gs ar wk pc mk ei as mc ve ao.
```

STEPS for USING the MULTIVARIATE RANGE RESTRICTION CORRECTION SPSS PROGRAM

The following steps describe how to use the Lawley multivariate correction for range restriction program developed by CNA in SPSS. The correction is necessary in order to fairly compare the (estimated) population validity coefficients of ASVAB composites. The unfair comparison without the correction is due to the military use of only a subset of tests (in a composite) with cutscore as an operational classification standard.

1) From the sample data file with all ASVAB tests, other predictor variables, and final school grade (FSG) (or any other criteria) compute any ASVAB composites of interest for the study (theoretical or empirically derived). A syntax file example is “Lawley_Compute_Composites.sps“.

2) Use the “Lawley_Cormat.sps” file to generate a correlation matrix from the sample data set (“samp_mat.sav”). You will have to name that matrix output file accordingly and save it after deleting all but one of the “N” lines.

3) Modify the “Lawley_Correction.sps” file to specify the list of variables as they are written in the “Lawley_Cormat.sps” file.

4) In the “Lawley_Correction.sps” file, change the existing drive and path to the sample matrix to the one you are working with (the file handle path).

5) Also, create a folder called “TEMP” in the drive for holding the intermediate files that are created during the running of the program.

6) The “CORPOP97.sav” file is provided as a screen shot in this appendix so you will have to manually (one time) overwrite the “samp_mat.sav” file to get the ASVAB PAY97 unrestricted matrix) called “CORPOP97.SAV”.

THE PROGRAM GENERALLY KNOWN AS THE LAWLEY CORRECTION FOR MULTIVARIATE RANGE RESTRICTION.

```
* PROGRAM ORIGIN AND HISTORY.

* The program was written in 1977 and has subsequently been used and improved by researchers at CNA and NPRST.
* It is based on the technique originally developed by Pearson [1] and later refined by Burt [2] and by Lawley [3].
* [1] Pearson, Karl, "On the Influence of Natural Selection on the Variability and Correlation of Organs,
*       Phil.Trans.Roy.Soc. London, A,(1902): 1-66.
* [2] Burt, Cyril. "Validating Tests for Personnel Selection," British Journal of Psychology 34 (1943): 1-19.
* [3] Lawley, D.N., "A Note on Kark Pearson's Selection Formulae,
*       Proc. Royal Soc. Edinburgh, Sec. A (1943), 62 Part I, pp 28-30.
* The program was originally written at CNA in APL and FORTRAN by Verna and Mifflin [4].
*[4] CNA, Research Contribution 336. "A Method to Correct Correlation Coefficients for the Effects of Multiple Curtailment,"
*       by Thomas L. Mifflin and Steven M. Verna, Unclassified, Aug 1977.
* The FORTRAN version was modified by Peter Stoloff of CNA circa 1985.
* The program was converted (CNA) to SPSS syntax in 1996 by Christine Baxter [5] with results verified by Cathy Hiatt.
* [5] CNA 96-0773, "SPSS Range Correction Program," by Christine Baxter, Unclassified, 10 May 1996.
* The program was further revised and commented by John H. Wolfe for NPRST, July, 2006.
* Instructions and sample data added by Janet Held of NPRST circa 2010.

* PROGRAM CODE AND COMMENTS.

* This version corrects MEANS as well as STANDARD DEVIATIONS and CORRELATIONS.

* NOTE TO USER: These 3 FILE HANDLE statements must be changed
     to specify the path names in your particular computer.

FILE HANDLE SAMPCORR   /NAME =  'C:\Documents\Lawley\samp_mat.sav'.
FILE HANDLE POPCORR    /NAME =   'C:\Documents\Lawley\CORPOP97.SAV'.
FILE HANDLE OUTCORR    /NAME =   'C:\Documents\Lawley\corr_mat.SAV' .

* NOTE TO USER: Change VARLIST to list all of your variables in samp_mat including the study's ASVAB composites.
* All ASVAB tests (treated as explicit selection variables) first followed by implicitly selected. The Assembling Objects (AO)
     test is treated by the Navy as an implicit (incidental) selection variable like final school grade (FSG)  because of changes
     since PAY97 that have improved the applicant AO distribution (a more suitably challenging test).
DEFINE VARLIST ()
               gs ar wk pc mk ei as mc ve ao vear vemk ar2mkgs fsg .

!ENDDEFINE .

* POPCORR contains the population values of the means, standard deviations, and correlations of the
     explicitly selected variables used for correcting for range restriction, and no other variables.
* SAMPCORR has the sample means, standard deviations, and correlations of the explicitly
     selected variables (in the same order as POPCORR) followed by the implicitly selected  variables.

* OUTCORR is the output file containing range-corrected means, standard deviations, and correlations.
```

```
Get FILE SAMPCORR.


SAVE OUTFILE ='C:\Documents\Lawley\TEMP\CORSMP.SAV'
 /COMPRESSED.
GET FILE = POPCORR.

DEFINE CRANGE().

*Convert population correlation to population covariance matrix.
MCONVERT
/MATRIX=IN(*)
/MATRIX=OUT(*).

SAVE OUTFILE = 'C:\Documents\Lawley\TEMP\COVPOP.SAV'
 /COMPRESSED.

*Read in sample correlation matrix.

GET FILE = 'C:\Documents\Lawley\TEMP\CORSMP.SAV'.
EXECUTE.

*Convert sample correlation to sample covariance matrix.
MCONVERT
/MATRIX=IN(*)
/MATRIX=OUT(*).

SAVE OUTFILE='C:\Documents\Lawley\TEMP\COVSMP.SAV'
/COMPRESSED.

* Begin matrix job.

MATRIX.

* Read in matrix file.
  MGET   /FILE = 'C:\Documents\Lawley\TEMP\COVPOP.SAV' .
  COMPUTE POPMNp= MN .
* Determine number of explicitly selected variables.
  COMPUTE p = NCOL(POPMNp) .
  COMPUTE Wpp = CV .
  RELEASE CV, NC,MN,SD.

  MGET /FILE = 'C:\Documents\Lawley\TEMP\COVSMP.SAV'  .

 COMPUTE SAMPNC=NC.
 COMPUTE SAMPMN=MN.
 COMPUTE SAMPSD=SD.
 COMPUTE V = CV.
```

```
*Determine size of sample matrix.
 COMPUTE n  = NROW(V).
 COMPUTE p1 = p+1 .

* Reference: Lord & Novick, Statistical Theories of Mental Test Scores.
* Reading, Mass.: Addison-Wesley, 1968, pages 146-147.

* Partition sample matrix V into p explicitly selected variables and n-p implicitly selected variables.

COMPUTE Vpp = V(1:p, 1:p).
COMPUTE Vpn_p = V(1:p, p1:n).

COMPUTE Vn_pp = TRANSPOS(Vpn_p).
COMPUTE Vn_pn_p = V(p1:n,p1:n).

COMPUTE B = Vn_pp*INV(Vpp) .
* B = Regression weights for predicting implicitly selected variables from explicitly selected variables
    Use of B simplifies Lawley's formulas.

* Calculate corrected partitions (W = corrected to population values).
COMPUTE Wn_pp =    B*Wpp.
COMPUTE Wn_pn_p = Vn_pn_p - B*TRANSPOS(Vn_pp) + Wn_pp* TRANSPOS(B) .

* Assemble corrected matrix from partitions.
COMPUTE W  =  {Wpp,TRANSPOS(Wn_pp);Wn_pp,Wn_pn_p}.

* Correct Means.
COMPUTE SMNp = SAMPMN(1:p) .
COMPUTE SMNn_p = SAMPMN(p1:n) .
* B0 is the row vector of constants in the regression equations.
COMPUTE B0 = SMNn_p - SMNp*TRANSPOS(B) .
COMPUTE POPMNn_p = B0 + POPMNp*TRANSPOS(B) .
COMPUTE POPMN = {POPMNp, POPMNn_p} .

* Save as a matrix file.

MSAVE POPMN
 /TYPE=MEAN
 /OUTFILE = 'C:\Documents\Lawley\TEMP\CORRECT.SAV'
 /VARIABLES= VARLIST .
 MSAVE SAMPNC
 /TYPE N.
MSAVE W
 /TYPE COV.

*End matrix job.

END MATRIX.
```

```
*Convert corrected covariance matrix to a correlation matrix.

GET FILE = 'C:\Documents\Lawley\TEMP\CORRECT.SAV' .
EXECUTE.

MCONVERT
 /MATRIX=IN(*)
 /MATRIX=OUT(*).

*Output corrected correlation matrix.
!ENDDEFINE.

CRANGE.

SAVE OUTFILE = OUTCORR  / COMPRESSED.

*NOTE TO USER: You can run DELTEMP.BAT to delete temporary files in C:\TEMP.
```

Screen Shot of PAY97 Matrix used for Navy ASVAB validation/standards studies

gss97pr	ars97pr	wks97pr	pcs97pr	mks97pr	eis97pr	ass97pr	mcs97pr	ves97pr	aos97pr
50.000	50.010	49.990	50.000	50.000	50.000	50.010	49.990	50.000	49.990
462039	462039	462039	462039	462039	462039	462039	462039	462039	462039
9.998	9.993	9.998	9.997	9.993	10.001	10.003	10.001	9.995	10.002
1.000	.721	.800	.717	.687	.700	.520	.684	.814	.568
.721	1.000	.671	.723	.798	.596	.423	.651	.733	.641
.800	.671	1.000	.764	.614	.611	.434	.583	.964	.509
.717	.723	.764	1.000	.675	.552	.345	.592	.906	.590
.687	.798	.614	.675	1.000	.484	.241	.551	.676	.598
.700	.596	.611	.552	.484	1.000	.724	.712	.624	.493
.520	.423	.434	.345	.241	.724	1.000	.671	.424	.380
.684	.651	.583	.592	.551	.712	.671	1.000	.622	.647
.814	.733	.964	.906	.676	.624	.424	.622	1.000	.573
.568	.641	.509	.590	.598	.493	.380	.647	.573	1.000

Appendix B
Generated Taylor-Russell .10 Base Rate Table

Taylor-Russell .10 Base Rate Table

	Selection Ratio																		
	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.9	0.95
Validity																			
0.00	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.1	0.1	0.1
0.05	0.12	0.12	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.10	0.10	0.10	0.10	0.1	0.1	0.1
0.10	0.14	0.13	0.13	0.13	0.12	0.12	0.12	0.12	0.12	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.1	0.1	0.1
0.15	0.16	0.15	0.15	0.14	0.14	0.13	0.13	0.13	0.12	0.12	0.12	0.12	0.11	0.11	0.11	0.11	0.11	0.1	0.1
0.20	0.19	0.17	0.16	0.15	0.15	0.14	0.14	0.13	0.13	0.13	0.12	0.12	0.12	0.12	0.11	0.11	0.11	0.11	0.1
0.25	0.22	0.19	0.18	0.17	0.16	0.15	0.15	0.14	0.14	0.13	0.13	0.13	0.12	0.12	0.12	0.11	0.11	0.11	0.1
0.30	0.25	0.22	0.20	0.19	0.18	0.17	0.16	0.15	0.15	0.14	0.14	0.13	0.13	0.12	0.12	0.12	0.11	0.11	0.1
0.35	0.28	0.24	0.22	0.20	0.19	0.18	0.17	0.16	0.15	0.15	0.14	0.14	0.13	0.13	0.12	0.12	0.11	0.11	0.1
0.40	0.31	0.27	0.24	0.22	0.20	0.19	0.18	0.17	0.16	0.15	0.15	0.14	0.14	0.13	0.12	0.12	0.11	0.11	0.1
0.45	0.35	0.29	0.26	0.24	0.22	0.20	0.19	0.18	0.17	0.16	0.15	0.15	0.14	0.13	0.13	0.12	0.12	0.11	0.1
0.50	0.39	0.32	0.29	0.26	0.24	0.22	0.20	0.19	0.18	0.17	0.16	0.15	0.14	0.13	0.13	0.12	0.12	0.11	0.11
0.55	0.43	0.36	0.31	0.28	0.25	0.23	0.21	0.20	0.19	0.17	0.16	0.15	0.14	0.14	0.13	0.12	0.12	0.11	0.11
0.60	0.48	0.39	0.34	0.30	0.27	0.25	0.22	0.21	0.19	0.18	0.17	0.16	0.15	0.14	0.13	0.12	0.12	0.11	0.11
0.65	0.53	0.43	0.37	0.32	0.29	0.26	0.24	0.22	0.20	0.18	0.17	0.16	0.15	0.14	0.13	0.12	0.12	0.11	0.11
0.70	0.58	0.47	0.40	0.34	0.31	0.27	0.25	0.22	0.21	0.19	0.18	0.16	0.15	0.14	0.13	0.12	0.12	0.11	0.11
0.75	0.64	0.51	0.43	0.37	0.32	0.29	0.26	0.23	0.21	0.19	0.18	0.16	0.15	0.14	0.13	0.12	0.12	0.11	0.11
0.80	0.71	0.56	0.47	0.40	0.34	0.30	0.27	0.24	0.22	0.20	0.18	0.17	0.15	0.14	0.13	0.12	0.12	0.11	0.11
0.85	0.78	0.62	0.51	0.43	0.36	0.31	0.28	0.25	0.22	0.20	0.18	0.17	0.15	0.14	0.13	0.12	0.12	0.11	0.11
0.90	0.86	0.69	0.56	0.46	0.38	0.33	0.28	0.25	0.22	0.20	0.18	0.17	0.15	0.14	0.13	0.13	0.12	0.11	0.11
0.95	0.95	0.78	0.61	0.49	0.40	0.33	0.29	0.25	0.22	0.20	0.18	0.17	0.15	0.14	0.13	0.13	0.12	0.11	0.11
1.00	1.00	1.00	0.67	0.50	0.40	0.33	0.29	0.25	0.22	0.20	0.18	0.17	0.15	0.14	0.13	0.13	0.12	0.11	0.11

Mathematically generated table by Mrs. Rebecca D. Hetter to reflect the Taylor-Russell .10 rate table.with shorter selection ratio intervals.

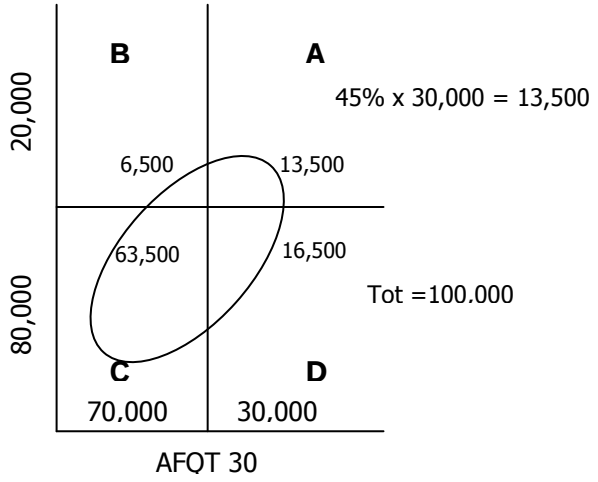
Taylor, H.C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology, 23,* 565-578.

Appendix C
Worksheet for Calculating Classification Decision Errors

Classification Decisions with Moderately Large Validity Coefficients, Stringent Selection (SR=.30), and a Series of High Population Success Rates (Base Rates)

Predictor	R_{xy}	BR=.20	BR=.25	BR=.30	BR=.35
ASVAB	.60	.45	.51	.55	.61
ASVAB + DLAB	.65	.48	.53	.57	.64
Expected Success Improvement		+3%	+2%	+2%	+3%

Note. Data were taken from Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology, 23*, 565-578.



BR = .20, R_{xy} = .60

A = 30,000 Qual x Success Rate (45%) = 13,500

D = 30,000 – 13,500 = 16,500

B = 20,000 – 13,500 = 6,500

C = 80,000 – 16,500 = 63,500,

or

C = [100,000 – (A+B+D)]

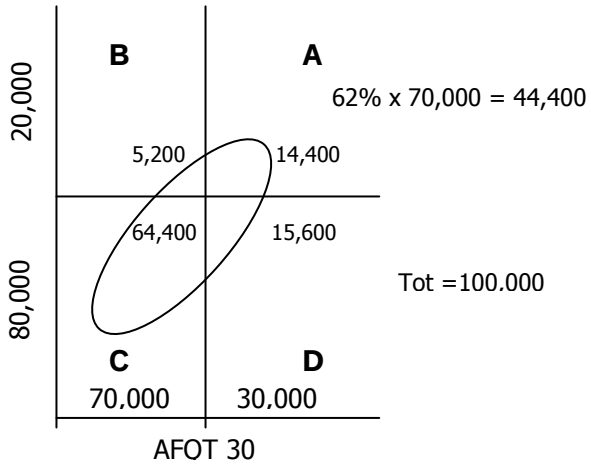
BR = .20, R_{xy} = .65

A = 30,000 Qual x Success Rate (48%) = 14,400

D = 30,000 – 14,400 = 15,600

B = 20,000 – 14,400 = 5,600

C = 80,000 – 15,600 = 64,400



Decisions	Correct+ A	Correct- C	Incorrect– B	Incorrect+ D
R_{xy} = .60	13.5%	63.5%	6.5%	16.5%
R_{xy} = .65	14.4%	64.4%	5.6%	15.6%
	.9% Improvement	**.9% Improvement**	**.9% Improvement**	**.9% Improvement**

BR = Base Rate – Population Success
SR = Selection Ratio – Resulting from Cutscore
R_{xy} = Validity coefficient – Applying to Population

Correct+ = Correct Acceptance
Correct- = Correct Rejection
Incorrect- = Incorrect Rejection
Incorrect+ = Incorrect Acceptance

Distribution List

AIR FORCE PERSONNEL CENTER, STRATEGIC RESEARCH AND ASSESSMENT BRANCH (HQ AFPC/DSYX)
AIR FORCE – FORCE MANAGEMENT POLICY DIRECTORATE, TRAINING AND EDUCATION REQUIRMENTS AND RESOURCES DIVISION (HQ AF/A1PT)
AIR UNIVERSITY LIBRARY
ARMY RESEARCH INSTITUTE LIBRARY
CENTER FOR NAVAL ANALYSES LIBRARY
CHIEF OF NAVAL PERSONNEL (OPNAV 132G, NAVY SELECTION AND CLASSIFICATION OFFICE)
DEFENSE MANPOWER DATA CENTER (CHIEF, PERSONNEL TESTING DIVISION)
MARINE CORPS MANPOWER AND RESERVE AFFAIRS, MANPOWER PLANS DIVISION, INTEGRATION AND ANALYSIS SECTION (SECTION HEAD)
MILITARY ACCESSION POLICY WORKING GROUP
NAVAL POSTGRADUATE SCHOOL DUDLEY KNOX LIBRARY
NAVAL RESEARCH LABORATORY RUTH HOOKER RESEARCH LIBRARY
NAVY PERSONNEL RESEARCH, STUDIES, AND TECHNOLOGY SPISHOCK LIBRARY
HQ USMEPCOM (DIRECTOR OF TESTING)
OFFICE OF NAVAL RESEARCH (CODE 34)
OFFICE OF THE UNDERSECRETARY OF DEFENSE (PERSONNEL & READINESS) (ASSISTANT DIRECTOR, ACCESSION POLICY)
USAF ACADEMY LIBRARY
US COAST GUARD ACADEMY LIBRARY