AFRL-RH-WP-TR-2020-0086

Air Force Personnel Center Best Practices Guide

Briefing Validation Results

U. Christean Kubisiak, Editor

Personnel Decisions Research Institutes, LLC

Scott B. Morris

August 2020

Interim Report

DISTRIBUTION A. Approved for public release.

AIR FORCE RESEARCH LABORATORY
711TH HUMAN PERFORMANCE WING
AIRMAN SYSTEMS DIRECTORATE
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

NOTICE AND SIGNATURE PAGE

AFRL-RH-WP-TR-2020-0086 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

//signature//
THOMAS R. CARRETTA
Work Unit Manager
Performance Optimization Branch
Airman Biosciences Division

//signature//
LOGAN A. WILIAMS
Airman Readiness Optimization CRA
Performance Optimization Branch
Airman Biosciences Division

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE (DD-MM-YY)
21-08-20

2. REPORT TYPE
Interim

3. DATES COVERED (From - To)
15-03-19 to 31-07-20

4. TITLE AND SUBTITLE
Air Force Personnel Center Best Practices Guide: Briefing Validation Results

5a. CONTRACT NUMBER
FA8650-14-D-6500, Task Order 0007

5b. GRANT NUMBER
Not applicable

5c. PROGRAM ELEMENT NUMBER
62202F

6. AUTHOR(S)
*Scott B. Morris and U. Christean Kubisiak

5d. PROJECT NUMBER
5329

5e. TASK NUMBER
09

5f. WORK UNIT NUMBER
H0SA (532909TC)_____

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

*PDRI, an SHL Company
1911 N. Fort Myer Drive
Suite 410
Arlington, VA 22209

8. PERFORMING ORGANIZATION REPORT NUMBER

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

Air Force Materiel Command
Air Force Research Laboratory
711th Human Performance Wing
Airman Systems Directorate
Airman Biosciences Division
Performance Optimization Branch
Wright-Patterson AFB, OH 45433

Air Force Personnel Center
Strategic Research and Analysis Branch
550 C St West, Ste. 45
JBSA-Randolph, TX 78150-4747

10. SPONSORING/MONITORING AGENCY ACRONYM(S)
711 HPW/RHBC

11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)
AFRL-RH-WP-TR- 2020-0086

12. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution A: Approved for public release. 88ABW-2020-3036, Cleared on 30 September 2020

13. SUPPLEMENTARY NOTES
Report contains color.

14. ABSTRACT

This series of reports is intended to consolidate experience and best practices the Air Force has accumulated in its selection and classification work. This report begins with an introduction to the Air Force Personnel Center Strategic Research and Assessment Branch (DSYX). It then goes on to describe best practices for presenting results of validation studies, including a review of relevant professional standards that guide reporting of research findings, discusses general and specific considerations for reporting criterion-related validity results, and covers strategies and techniques for effectively communicating validity evidence.

15. SUBJECT TERMS
Selection, Classification, Air Force, Validation, Criterion-related Validity

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT:**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON (Monitor)**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Thomas R. Carretta
Unclassified	Unclassified	Unclassified	SAR	95	**19b. TELEPHONE NUMBER (Include Area Code)** N/A

Table of Contents

LISTOF FIGURES

LISTOF TABLES

FOREWORD

This report is one of a series that compile the best of the experience, wisdom and tools that the Air Force has accumulated in its selection and classification work, and best practices from industry and academia. These reports draw upon the experiences of the Air Force Personnel Center/Strategic Research and Assessment branch (AFPC/DSYX) and leading researchers and practitioners in the field of Industrial/Organizational Psychology to provide guides to cover a variety of topics. Each begins with a section describing AFPC/DSYX and the background of their research to provide context for the series. This report addresses best practices in reporting and briefing results of data driven research.

EXECUTIVE SUMMARY

This series of reports is intended to consolidate the experience, wisdom, and tools that the Air Force has accumulated in its selection and classification work, and to blend these with best practice recommendations from industry. The reports cover a wide variety of material, including chapters on test development and validation, selection/classification model development, reporting/briefing results, and ethical and legal considerations. The goal is to ensure consistency as AFPC/DSYX continues to develop assessments and refine selection and classification models for a large number of Air Force career fields

We begin with an introduction to AFPC/DSYX. The background and history are covered, describing how the Air Force Human Resources Laboratory and its elimination left a need for providing research in human capital management. That was resolved in 2010 with funding to create DSYX which is intended to review, evaluate, develop, validate, and manage personnel programs to improve recruiting, selection, classification, and utilization of military personnel. The chapter describes how DSYX contributes to strategic human capital management, tools it makes available for testing, experience and expertise it provides, and looks ahead to the future and how DSYX can build on its capabilities.

The body of this report describes best practices for presenting results of validation studies. A guiding principal of employee selection continues to be the use of empirically based decision making. Selection systems involve collecting data on the psychometric quality, job relevance and predictive accuracy of assessments, and using those results to make informed decisions about whether assessments can be successfully implemented.

The utility of those assessments depends both on the quality of the data used, choices in the design and implementation of the assessments, careful handling and cleaning of the data, and proper interpretation of the results. Therefore, it is critical that research reports be transparent with regard to the methodology employed, as well as research reports and findings.

Proper deployment of selection techniques requires a high degree of technical sophistication to properly use and interpret results from complex psychometric and statistical analyses. Because of this, the selection expert, in communicating this information must execute a careful balancing act, maintaining the precision and transparency demanded by professional standards, while simultaneously conveying the essence of findings to a non-technical audience.

This report provides a review of relevant professional standards that exist to guide reporting of research findings, both at a broad level and specific to criterion related validation. It then discusses general considerations for reporting criterion-related validity results, and specific guidance for information that should be minimally provided. Finally, it covers strategies and techniques for effectively communicating validity evidence in a way that retains the required technical information, but making it accessible to audiences.

Introduction to the Air Force Personnel Center Strategic Research and Assessment Branch (AFPC/DSYX)

Background/History

Human Capital Management Mandates. The Air Force Policy Directive, AFPD 36-XX, Air Force Personnel Assessment Program, raised the bar for validation of Air Force operations affecting human capital management. The policy directive laid out Air Staff-defined objectives in support of both 1) DoD initiatives, such as the Testing Modernization Program, supported by major influxes of research and development funding and 2) the Human Capital Annex of the Air Force Strategic Personnel Plan (moving ahead with several active Air Force-level working groups). The Air Force's way forward in support of these flow-down mandates included both the objectives and the scope of this initiative:

- Establish processes to apply scientific analysis and technology in support of recognized best practices to support personnel assessment. The goal of the Air Force Personnel Assessment Program is to support effective force management by ensuring that the right persons having the right aptitudes, characteristics, skills, and abilities are identified and accessed into the Air Force, are properly trained, and then optimally utilized to support the Air Force mission.
- The Air Force Personnel Assessment Program includes, but is not limited to, selection and classification, promotion, and proficiency assessment; and survey capability for assessing attitudes and opinions, job performance, and Air Force Specialty (AFS) requirements and characteristics.

Air Force Human Resources Laboratory

In 1968, the broad personnel research efforts (e.g., manpower, personnel, training) from various programs across the Air Force were consolidated into the Air Force Human Resources Laboratory (AFHRL). The name "Air Force Human Resources Laboratory" was only used as the official designation for the combined program from 1968 to 1991. However, it was the name used for the longest period of time and is the one that has the greatest familiarity to professionals, in and out of the government, with an interest in military psychology. The antecedents of AFHRL can be traced to the Psychological Research Units of the Aviation Psychology Program in the Army Air Corps during World War II. After the Air Force became a separate service in 1947, AFHRL was called the Human Resources Research Center (1949-1953), Personnel and Training Center (1954-1958), Personnel Laboratory (1958-1962), and Personnel Research Laboratory (1962-1968). In 1991, the name Air Force Human Resources Laboratory was retired and the mission was absorbed by successor organizational units within the Armstrong Laboratory (1991-1996) and the Air Force Research Laboratory (1997-1999). In 1999, the personnel research function in the Air Force (Manpower and Personnel Research Division) was eliminated, leaving no organizational entity for research in the domains of personnel selection and classification.

The Rise of the Strategic Research and Assessment Branch (AFPC/DSYX)

The need for research in strategic human capital management within the Air Force did not end with the elimination of AFHRL funding. After the elimination of AFHRL, minimal funding was provided to manage testing-related contracts and provide basic support for operational testing programs. In 2010, additional funding was provided to create the AFPC/DSYX program and several billets were created to continue the work that ended with the elimination of AFHRL in 1999.

AFPC/DSYX Program Overview

With the additional funding, the DSYX program was tasked to review, evaluate, develop, validate, and manage personnel programs to improve recruiting, selection, classification, and utilization of military personnel. The current responsibilities of DSYX include Air Force- and Department of Defense-related testing programs, research and analysis, and development and validation of new assessment processes and measures. The DSYX program now develops person-job match screening processes to support optimal personnel utilization for the entire personnel life cycle including pre-recruiter job exploration (e.g., interest inventories, realistic job previews); applicant assessment, screening, and classification of recruits (e.g., cognitive, personality, psychomotor, occupation-specific assessment of skills), retraining, and specialized assignments.

The DSYX program also helps maintain a mission-ready force by managing Air Force Specialty Code (AFSC) structures using scientific standards to establish desirable and mandatory occupational entry requirements and adjust occupational structures to optimize training investment, career progression, utilization, and retention for total force integration. Thus, the ultimate purpose of the DSYX program is to provide: 1) consultation to program managers and Air Force leadership on selection and classification issues, 2) development, revision, and validation of personnel tests, 3) technical oversight of the operational testing program, and 4) management of contracts in support of personnel-related research.

AFPC/DSYX Organizational Structure

The DSYX branch is now embedded within the Air Force Personnel Center (AFPC) Directorate of Staff. As previously mentioned, while no longer supported by a multitude of scientists and psychologists, DSYX provides an array of services and tools similar to AFHRL. The current structure of DSYX includes the branch chief, a program manager, seven personnel research psychologists, and two research assistants. While many of the tasks assigned to DSYX and much of the funding to accomplish them come from Air Staff (A1) and Air Force Testing Policy (A1PT), DSYX is officially under the command of AFPC.

Synergistic Relationships

The AFPC Promotions, Evaluations, and Recognition Branch (AFPC/DP3SP) manages the operational personnel testing program. Thus, while DSYX has the responsibility of developing and validating the tests within the personnel testing program, the operational responsibility of military testing resides with DP3SP. The one current exception is the Pilot Candidate Selection

Method (PCSM; described later in this report) which has been developed, validated, and operationally maintained by DSYX.

The Air Force Recruiting Service (AFRS) Operations Division's Analysis Branch (AFRS/RSOA) supports DSYX through participation in the regular working group conference calls with AF/A1PT and DSYX, pre-accession process advisories, data collection facilitation, collaborative ad hoc analyses, and unrestricted access to relevant operational data. AFRS/RSOA also assists in implementation of new selection and classification assessment measures and processes. These activities are consistent with an operational mandate to support improving selection and classification systems (tests and processes) to optimize recruiting efficiency for Air Force Officer and Enlisted accessions while continuously adapting to changing population characteristics, training dynamics/criteria, and needs of the Air Force.

The AFPC/DSYX Contribution to Human Capital Management and Strategic Human Resources Management through Mission Alignment

DSYX makes contributions to the Air Staff by following the mission as tasked by AFMAN 36-2664:

- Provide technical guidance to and consult with AF/A1PT in identifying and overseeing strategic human resource capital initiatives.
- Support human capital studies and research to support decision-making involving recruiting, selection, classification, promotion, utilization, and retention.
- Coordinate changes to Air Force Officer Classification Directories (AFOCD) and Air Force Enlisted Classification Directories (AFECD).
- Support revision and validation of the Air Force Officer Qualifying Test (AFOQT), the Pilot Candidate Selection Method (PCSM), and the Test of Basic Aviation Skills (TBAS).
- Conduct development, validation, and revision of tests and assessments.
- Evaluate enlistment and commissioning standards.
- Provide technical oversight of operational selection, classification, utilization, promotion, and proficiency testing and assessments to ensure they meet professional and legal standards.
- Technically review requests to develop/implement new tests/assessments.
- Manage the Applied Performance and Assessment Testing Center at Lackland AFB.

DSYX makes contributions to the Air Force Personnel Center by following the mission as tasked by AFPC Mission Directive 37, 2003 [1-up]:

- Manage and operate Air Force military personnel data and information systems, execute policies that govern active duty accessions, testing, classification, assignments, personnel record systems, and personnel assessment.
- Manage and operate Air Force civilian personnel data and information systems and personnel assessment programs.

The DSYX Testing Toolbox

General Ability/Aptitude Tests

Air Force Officer Qualifying Test (AFOQT). The (AFOQT is used to help select candidates for officer commissioning programs and to classify commissioned officers into utilization specialties such as manned aircraft pilot, RPA pilot combat system operators, air battle manager, or technical. Air Force Officer Qualifying Test scores are also used as a quality metric in the integrated officer classification model. The AFOQT is available in two versions (Form T1 and T2). Each version consists of 12 subtests. Subtests are used to compute one or more of the five aptitude composites. Scores on the subtests relate to performance in certain types of training. AFOQT composite scores are reported in percentiles.

Armed Services Vocational Aptitude Battery (ASVAB). The ASVAB evaluates specific aptitude areas and provides a percentile score related to requirements for selecting and classifying individuals for the Armed Services. There are two ASVAB testing programs—Student and Enlistment. The Student Testing Program applies to ASVAB testing in educational institutions such as high schools and vocational trade schools. The Enlistment Testing Program applies to Armed Services Vocational Battery testing in authorized accessions testing facilities such as Military Entrance Processing Stations (MEPS) and Military Entrance Test Sites (METS). The Army is the executive agent for the overall ASVAB Testing Program. The Defense Personnel Assessment Center in the Office of People Analytics is the executive agent for the ASVAB. The Air Force computes four training classification composite scores for the ASVAB: Mechanical (M), Administrative (A), General (G), and Electronics (E). These scores are predictive of training success in a variety of military occupations.

Electronic Data Processing Test (EDPT). The EDPT evaluates the basic ability to complete formal courses for programming electronic data processing equipment. The EDPT is a multiple-choice test that contains measures of verbal ability, symbolic reasoning, and arithmetic reasoning. It is used to screen and select Airmen for career fields requiring this ability. It is available by paper-and-pencil and electronically on the Personnel Testing Station[1] platform.

Vocational Interests

Air Force Work Interest Navigator (AF-WIN). The AF-WIN is an internet-delivered interest inventory that matches examinees' interests on the dimensions of functional communities, job contexts, and work activities to AFSC job profile markers to identify their "best fit" Air Force Specialties. It takes 15-20 minutes to complete with the examinee indicating level of interest on a 5-point scale for 52 items. There is a version of the AF-WIN for enlisted AFSCs and two officer versions. One officer version is designed for use at the beginning of college to help examinees plan their curriculum to include coursework required for particular AFSCs. The second version is

[1] The Personnel Testing Station was formerly called the Test of Basic Aviation Skills test station.

for use closer to commissioning when finalizing the AFSC assigned to a cadet upon commissioning.

Personality

Tailored Adaptive Personality Assessment System (TAPAS). The TAPAS uses a trait taxonomy that assesses facets of the Big Five personality factors using a multidimensional pairwise preference (MDPP) format. The assessment requires about 30 minutes to complete. It is completed by all new recruits at the Military Entrance Processing Station at the same time they complete the Armed Services Vocational Aptitude Battery. It is also administered on the Personnel Testing Station platform for selected retraining AFSCs.

Self-Description Inventory (SDI). The SDI was first implemented on AFOQT Form S as a 220 item, trait-based personality assessment of the Big Five personality domains and two Air Force related scales (Team Orientation and Service Orientation). Factor analyses of SDI item content revealed broad six domains encompassing the Big Five domains plus Machiavellianism, with subsequent factor analyses of domain content revealing a total of 20 narrower trait facets. The AFOQT Form T version of the SDI contains 240 items that assess the Big Five personality domains and Machiavellianism and 30 underlying facets.

Although the SDI was initially developed for the U.S. Air Force, a collaborative initiative with allied forces led to adaptations of the SDI for research purposes in the militaries of Canada, United Kingdom, New Zealand, and Australia.

Miscellaneous/Specialty

Test of Basic Aviation Skills (TBAS). The TBAS is a battery of cognitive, multi-tasking, and psychomotor subtests administered on a computer test station. Examinees are required to respond to computerized tasks using a keypad, joysticks, and foot pedals. The TBAS includes subtests measuring psychomotor coordination, cognitive abilities, and multi-tasking capabilities. A pilot candidate's AFOQT Pilot composite score (or, where applicable, Enlisted Pilot Qualifying Test [EPQT] score) and Federal Aviation Administration certified flying hours are combined with the TBAS measurements to formulate a PCSM score. Manned aircraft Pilot and RPA pilot selection boards receive each candidate's PCSM composite score on a percentile scale of 1 to 99. PCSM assists pilot selection boards to select candidates most likely to successfully complete undergraduate pilot training.

Air Traffic Scenarios Test (ATST). The Air Traffic Scenarios Test is part of the classification screening process for candidates for the enlisted Air Traffic Control (ATC) AFSC. The Air Traffic Scenarios Test consists of simulated Air Traffic Control scenarios where the examinee is scored on how effectively they manage the departure, landing, tracking, etc. of aircraft with minimal safety violations. The test is administered on the TBAS testing platform and takes about an hour to complete.

Multi-Tasking Test (MTT). The Multi-Tasking Test measures the ability to shift attention from one task to another over a short period of time. The test includes four component tasks: Math, Visual, Memory, and Listening. In the math task, participants add three-digit numbers. In the memorization task, a list of letters is initially presented and then disappears; after a delay, a probe letter is presented and participants indicate whether or not the probe letter was included in the list. In the listening task, participants respond with a mouse click when they hear a high-pitched tone and ignore a low-pitched tone. Finally, in the visual monitoring task, a needle moves from right to left across a display resembling a fuel gauge and the goal is to reset the needle when it nears the end of the display. The test is administered on the PTS testing platform and takes about 45 minutes to complete.

The DSYX Expertise and Resources Toolbox

Staff Expertise

- Test Development/Validation – Professionals in the DSYX team have decades of experience in item writing, item selection, scale development, test development, and test validation. Current DSYX team members have experience developing DoD tests such as AFOQT, ASVAB, SDI, and AF-WIN. In addition, the team has experience in commercial test development including globally-recognized tests such as the Wechsler scales, the Beck inventories, and employee selection tests such as the Watson-Glaser Critical Thinking Appraisal and the Bennett Mechanical Comprehension Test.
- Predictive Model Development/Validation – Numerous occupational-specific predictive models have been developed by DSYX using pre- and post-accession tests. Numerous empirical and regression-based formulas to predict important performance-based outcomes have now been operationalized for selection and classification purposes.
- Job/Occupational Analysis – DSYX members have extensive expertise in job/occupational analysis to include task, trait, and competency analysis. The results of numerous DSYX-based job analysis studies are now used in developing predictive models, responding to career field inquiries, and setting standards for classification (e.g., based on ASVAB profiles).
- Vocational Interest – DSYX personnel have enlisted- and officer-level vocational interest inventories. The tools developed by DSYX have moved beyond traditional, generic vocational interest inventories and are specific to Air Force occupational specialties. The inventories provide information on the ideal match between a potential recruit and an occupational specialty and provide guidance to the examinee regarding the cognitive and physical requirement for the job.
- Job Satisfaction – DSYX personnel have conducted studies of job satisfaction using USAF Occupational Analysis (OA) data and internally-developed surveys to determine if DSYX tests and/or predictive models are contributing to improved satisfaction.
- Structured Interviews – DSYX has worked with USAF career fields to create structured interviews, structured interview guides, and video-based instructions for conducting valid structured interviews.
- Ethics/Integrity – DSYX staff members have extensive experience in ethical behavior, integrity, and counterproductive behavior. DSYX has developed integrity tests and valid tests designed to detect the propensity to engage in counterproductive behavior.
- Realistic Job Preview Creation – DSYX staff members have extensive expertise in developing realistic job preview videos based on SME input video-based interviews.

- Leadership – DSYX staff members have extensive expertise in assessing theories/models of leadership competencies and in the evaluation of leadership potential to help senior leaders attract, develop, and retain talent to effectively and efficiently accomplish mission requirements. The expertise encompasses experiences gained through work in academia, private industry, and military/government, which aid in providing customers with valuable tools, analysis, and innovative insights designed to improve organizational performance.

Contractor Expertise

Consulting Firms. DSYX has had the opportunity work with the most well-known consulting firms in industrial and organization psychology and government research. In addition, DSYX has been able to contract out some work to the most recognized experts in their respective fields, including former presidents of the Society of Industrial and Organization Psychology (SIOP) and leading authors in academia and cutting-edge commercial innovation.

Forward Looking: The Future of DSYX

Increased Effort to have DSYX Expertise, Services, and Interventions Recognized throughout the Air Force

Recent efforts by DSYX have improved the visibility of the branch throughout the Air Force. Specifically, efforts to educate Career Field Managers (CFMs) on the tools and services provided by DSYX have resulted in operational Predictive Success Models for numerous career fields and expansion of the use of existing tests for selection and classification purposes. In addition, updated internal marketing materials (e.g., slide decks, tri-fold brochures) are being prepared to provide additional exposure for the beneficial offerings of DSYX. Finally, high-profile attention to quality products such as AF-WIN are providing additional recognition for how DSYX can provide high-quality and cost-effective services to the Air Force. Additional efforts will need to be expended in this area in order for DSYX to continue to thrive as a valuable internal asset.

Improved Technology

Recent and future advances in available technology will provide DSYX with the capability to provide services and tools in a more efficient manner. Examples include item-banking, a combined test-development and test-delivery platform, and even sophisticated tools such as text analysis.

Improved Access to Data

Current processes to procure and process necessary data (e.g., test scores, training grades) are somewhat inefficient and hinder the efficiency and effectiveness of the branch. Future enhancements are being vetted and implemented to automate and streamline the process. This will allow DSYX to provide real-time decision support to internal clients and will improve the speed in which DSYX can build the tests and tools required for effective selection and classification purposes.

Exiting the Operational Testing Domain

DSYX historically has been involved in many aspects of operational testing (e.g., test delivery, scoring, coding) which limits the time and resources available to devote to true mission-specific activities. Current efforts are being conducted to ensure a more efficient hand-off from DSYX to the operational entities after successful development of tests and selection/classification tools.

Repeatable and Scalable Processes

DSYX is currently striving to develop repeatable (e.g., consistent analyses, similar technical report templates) and scalable analyses and processes (e.g., processes that can be applied to large and small requests throughout the Air Force). This Guide is one small step in achieving that goal.

1.0 BEST PRACTICES FOR BRIEFING VALIDATION RESULTS

Scott B. Morris

1.1 Introduction

Evidence-based decision making has long been a guiding principal of employee selection. The design of a selection system involves collecting data on the psychometric quality, job relevance and predictive accuracy of assessments, and using the results to make informed decisions about whether assessments should be adopted, modified, or eliminated.

The usefulness of empirical results depends greatly on the quality of the data used in an analysis. Choices in the design and implementation of research, cleaning of data, and analysis can substantially impact the legitimacy and generalizability of results, and it is therefore critical that research reports be transparent with regard to the methodology employed. Responsible research practice also requires full transparency of research reports and findings.

Selection research requires a high degree of technical sophistication in order to properly use and interpret results from complex psychometric and statistical analyses. At the same time, the selection expert must be able to communicate this information to policy-makers and other stakeholders who often have limited technical backgrounds. Effectively communicating the results of selection research requires a careful balancing act, maintaining the precision and transparency demanded by professional standards, while simultaneously conveying the essence of findings to a non-technical audience.

This report will start with a review of relevant professional standards, followed by specific recommendations for reporting criterion-related validity results. Several strategies for effectively communicating validity evidence will then be discussed.

1.2 Professional Standards for Reporting Validity Evidence

Several professional standards exist to guide reporting of research findings, including both broad standards for scientific work as well as guidelines specific to criterion-related validation. The following section provides a brief overview of several relevant standards, followed by an integrated set of recommendations for writing validation reports.

1.2.1. Sources of Professional Standards

Standards for Educational and Psychological Testing
Developed through a collaboration between the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (2014), the *Standards* provide broad guidance for the development and evaluation of tests, including recommendations for demonstrating the psychometric quality of measures (reliability and validity), test fairness, as well as recommendations for test administration and proper use of test scores. Relevant standards for validation research are found in Chapter 1, *Validity*, and Chapter 11, *Workplace Testing and Credentialing*.

Principles for the Validation and Use of Personnel Selection Procedures

Developed by the Society for Industrial and Organizational Psychology (2018), the *Principles* specify established scientific and professional practices related to the choice, development, evaluation, and use of personnel selection procedures designed to measure constructs related to work behavior. Recommendations for preparation of a technical validation report are provided on pp. 33-35.

Uniform Guidelines on Employee Selection Procedures (1978)

The *Uniform Guidelines* were developed to establish a uniform set of standards by which federal enforcement agencies can evaluate employee selection procedures in the context of prohibiting employment discrimination. These guidelines have been adopted by the Equal Employment Opportunity Commission, Department of Labor, Department of Justice, and Civil Service Commission.

While the Uniform Guidelines were informed by scientific principles and professional practices, they represent regulatory requirements imposed by government agencies, and are not a statement of scientific or professional principals. As such, the Uniform Guidelines tend to be more specific and directive in comparison to the broad principles stated in other guidelines. Unlike other professional standards, which have been regularly revised to represent advances in scientific understanding, the Uniform Guidelines have not been updated since 1978. Testing professionals have criticized the Uniform Guidelines for failing to keep pace with advances in validation methodology (Jeanneret & Zedeck, 2010; McDaniel, Kepes & Banks, 2011).

Journal Article Reporting Standards

The American Psychological Association (Appelbaum, Cooper, Kline, Mayo-Wilson, Nezu & Rao, 2018) published the *Journal Article Reporting Standards (JARS)* in an effort to promote rigor and transparency in scientific work by providing recommendations to authors and journal editors regarding the information to be included in research reports. Because the JARS are intended to be applicable to a broad spectrum of research areas, the recommendations tend to be very general, and do not specifically target reporting of validation research. JARS is available online at https://apastyle.apa.org/jars/

American Statistical Association Statement on Statistical Significance and P-Values

Statistical significance testing is a central component of most quantitative research, including validation studies. Despite its widespread use, the practice of significance testing has been widely criticized due to the common misinterpretation and misuse of significance tests (e.g., Schmidt, 1996). Through this statement (Wasserstein & Lazar, 2016), the American Statistical Association sought to clarify several accepted principles regarding the proper use and interpretation of significance tests. Due to the specific and technical nature of the recommendations, these guidelines are discussed in a later section of the report that focuses on significance testing.

1.2.2. Summary of Scientific and Professional Standards for Presenting Validation Research

Appendix A provides a summary of scientific and professional standards relevant to research on criterion-related validity in employment contexts. Standards related to other research contexts, other types of validity evidence (content, construct), and other aspects of selection systems (e.g., test fairness) are not included. The following guidelines were derived by integrating recommendations across the multiple professional and scientific standards.

A research report documenting a criterion-related validation study should include the following information:

1. Describe the intended uses of the selection system and the justification for the assessment procedures. What job-related qualifications is the test intended to assess? How will scores be used to make selection decisions (e.g., ranking vs. cutoff, multiple hurdles, combining multiple predictors)? The validation study should be designed to evaluate the intended interpretation and use of test scores.
2. Fully describe all variables included in the validation study:
    a. Identify all predictor and criterion variables, as well as any control variables used in the analyses. Identify all variables examined, even those that were not retained in the final analyses.
    b. Provide information on the psychometric properties of all measures (reliability, construct validity).
    c. Fully describe the test or assessment procedures being validated, including a description of the predictor constructs, test content, response process, and scoring procedures. When scoring involves judgment, information about rater selection, training, and scoring criteria should be provided. Include references for additional information on test development or technical manuals.
    d. Fully describe criterion measures and the data collection process, including steps taken to enhance the quality of measurements (e.g., training, use of multiple raters, etc.).
    e. Provide evidence that criterion measures reflect important work behaviors or work outcomes (e.g., linking measures to job analysis).
3. Describe the sample of individuals included in the validation study.
    a. Report the demographic composition (e.g., race, ethnicity, sex, age) and relevant work-related characteristics (e.g., applicants vs. incumbents, positions held, work experience). Discuss the representativeness of the sample in relation to the target population.
    b. Describe the procedures for recruiting participants and report the percent of recruited participants who were included in the final sample. Describe any inclusion/exclusion criteria and number of candidates excluded for each specific reason.
    c. Report the sample size (separately for each analysis if different) and power analysis.
4. Data collection procedures should be described in enough detail that testing professionals can evaluate the appropriateness of conclusions and make independent recommendations. Sufficient detail should be provided that a testing professional competent in personnel selection could replicate the study. Any research design factors that might impact the interpretation or generalizability of findings (e.g., low reliability, criterion contamination, range restriction, missing data, etc.) should be clearly stated.
5. Describe any data diagnostics conducted, including examination of data distributions or identification of statistical outliers. Specify any modification of the data resulting from

data cleaning, including variable transformations, exclusion of scores or participants, etc. Report sensitivity analysis on the impact of data cleaning procedures.

6. Provide details on the statistical analyses conducted:
    a. Describe the extent of missing data and how missing values were handled (e.g., case-wise or pairwise deletion, imputation).
    b. Provide descriptive statistics (e.g., frequencies, means, standard deviations) on all variables, both for the full sample and any relevant subgroups. For multivariate analyses, provide the full correlation or covariance matrix of variables.
    c. Report any problems with statistical assumptions (e.g., non-normal distributions, unequal error variance) that might impact the validity of findings.
    d. When reporting statistical significance tests, provide the test statistic, degrees of freedom (if appropriate) and *p*-value. Report the results of all tests conducted, not only those that were statistically significant.
    e. Report measures of effect sizes and confidence intervals where appropriate.
    f. When conducting complex analyses (e.g., structural equation modeling, hierarchical linear modeling, etc.), specify the software used in the analysis, and any relevant options used in the analysis (e.g., estimation method).
7. The presentation of results should strive to accurately and comprehensively portray the findings.
    a. A clear distinction should be made between the primary analysis, planned secondary analyses, and exploratory analyses.
    b. Report results for all variables examined, not just those that were statistically significant. When multiple analyses are conducted before identifying a final model, those preliminary analyses should be briefly described (although not in as much detail as the final model).
    c. When statistical adjustments are made (e.g., corrections for range restriction or measurement error), both adjusted and unadjusted coefficients should be reported, as well as the specific procedure used and all statistics used in the adjustment.
8. Conclusions and recommendations should be explicitly linked to study findings.
    a. Research findings that qualify conclusions or limit generalizability should be discussed.
    b. Efforts should be made to help readers correctly interpret results and avoid common misinterpretations.
    c. Research design factors or data analysis choices that potentially limit the validity or generalizability of findings should be acknowledged, and the potential impact on the findings discussed.

1.2.3. Targeting Presentation to the Audience

Professional standards for reporting research seek to promote scientific rigor, transparency, and replicability. At the same time, comprehensive reporting requirements demand time and space that is not available in all reporting formats. It is simply not feasible to include all required elements in concise presentation formats, such as briefing reports or oral presentations. Additionally, the depth of technical information could be overwhelming to a non-technical audience. For individuals with limited background in statistics, psychometrics, and validation research, the central findings could easily be lost in the excess of statistical details.

An important component of effective communication is understanding the audience (Aguinis, Werner, Abbott, Angert, Park & Kohlhausen, 2010). Different types of stakeholder tend to focus on distinct outcomes. While selection researchers might be interested in the technical quality of assessments (e.g., reliability and validity), hiring managers will be most concerned with the ability to meet staffing needs and performance standards. Others will focus on outcomes directly relevant to specific strategic objectives, such as increasing the representation of under-represented minority groups. Research reports will be most effective when framed in terms of the objectives valued by the intended audience.

It is useful to distinguish three levels of detail for research reports. A *technical report* is a written document intended to provide a comprehensive summary of the validation study and findings. Technical reports should strive for transparency and should generally include all elements recommended by scientific and professional standards. The primary audience will be selection researchers and therefore it is appropriate to include highly technical information. All findings should be fully reported, including relevant descriptive statistics, complete results of statistical analyses, and supplementing analyses.

A *briefing report* is a one to two page written document offering a concise summary of key findings and recommendations. The briefing report should be accessible to a non-technical audience.

A *briefing presentation* is a formal oral presentation of findings. Key information is presented in a concise format through a slide deck and supplemented with explanations delivered by the presenter. Like the briefing report, briefing presentations should be accessible to non-technical audiences. Additionally, presenters often operate under strict time limits, requiring selective presentation of findings. Finally, excessive text and data on slides can compete for the audience's attention rather than supporting the speaker's message and can lead to information overload (Mayer & Moreno, 2003).

In the following sections, recommendations for reporting validation results will initially be presented in the comprehensive format appropriate for a technical report, followed by more concise formats typical of briefing reports and presentations.

1.3 General Considerations in Presenting Statistical Results

A variety of statistical analyses are used in validation research, each with its own statistical indices and reporting traditions. Common to all analyses is the need to convey two basic types of information: an estimate of effect size and an indication of the precision of that estimate.

1.3.1. Effect Size

An effect size is a statistic representing the magnitude of some phenomenon (Kelley & Preacher, 2012). In the context of a validation study, effect size is a quantitative index of the strength and direction of the relationship between predictor and criterion variables. Effect size statistics help the researcher to understand the practical significance of findings; that is, how useful a predictor will be in identifying successful employees.

In order to be useful as a measure of practical significance, an effect size should be reported on a scale that has an easily understood metric. Some statistics, such as percentages, have a metric that is widely understood by most audiences (Kuncel & Rigdon, 2013). For example, most audiences would intuitively understand a report that 70 percent (%) of employees hired using a selection system would be successful. However, for many statistics, the metric is less clear, especially to those who do not have extensive experience with statistics to provide a common frame of reference (e.g., is a validity coefficient of 0.2 good or bad). In such cases, it is useful to provide benchmarks for representing different levels of practical significance (e.g., small, medium and large effect; Cohen, 1988). Such benchmarks have been criticized for being somewhat arbitrary and because the standards for usefulness vary depending on the specific research context (Bosco, Aguinis, Singh, Field & Pierce, 2015). Nevertheless, benchmarks provide a common frame of reference that helps the reader interpret the strength of relationships.

Measures of effect size are often standardized, resulting in a consistent metric regardless of the scale of the variables involved. Because the range of scores can vary substantially across different measures, results reported in terms of raw scores are often uninterpretable to those not familiar with a particular measure.

The use of standardized effect size statistics (e.g., correlation coefficients or standardized mean differences) is useful to researchers because it provides a consistent and readily interpretable index of the magnitude of results. However, the abstract nature of these statistics, which allows them to be generalized across settings, also makes them more difficult to interpret from the perspective of non-researchers. Aguinis et al. (2010) note the distinction between effect size statistics and information about practical significance. Information about whether a result is of practical significance is necessarily specific to a particular context and problem. Thus, when reporting the practical significance of results, efforts should be made to contextualize and interpret results in light of the priorities of the stakeholder, and should be communicated in terms of how the client views the problem. Several strategies to communicate the practical significance of validation results are discussed in a later section of the report.

1.3.2. Statistical Inference

In addition to the effect size, reports of statistical findings should indicate the degree of precision that the effect size estimates. In understanding the concept of precision, it is useful to distinguish between a population parameter and a sample statistic. The parameter is the theoretical value that would be obtained if data were available on the entire population of interest, while the statistic is the actual value computed from the data available. Due to the fact that the statistic was estimated using a limited sample, the value of a statistic will differ from the parameter. In other words, if you repeated a validation study, you would not obtain exactly the same estimate of the relationship. These differences, due to the limited sample of individuals included in a validation study, are called ‘sampling errors’. These sampling errors tend to be larger when sample size is small and shrink as sample size increases.

The sample statistic provides a point estimate (our best guess) of the population parameter. However, we also need to report the degree of uncertainty in that estimate. There are several ways to quantify the uncertainty resulting from sampling error. The standard error of the statistic represents the standard deviation of sampling errors, or the typical size of sampling error.

The standard error can be used to construct a confidence interval around the sample estimate. To construct a confidence interval, we choose a confidence level (CL = 95% is common) and build an interval around the point estimate. Say we compute a correlation of .3 with a 95% confidence interval of [.2, .4]. We know that the population correlation is probably not .3 exactly and the confidence interval indicates how far off our estimate might be. Specifically, if we construct a large number of intervals in this manner, 95% of the intervals will contain the population value.

Another common way to represent uncertainty due to sampling error is to conduct a test of statistical significance. We start with a null hypothesis, generally that the parameter of interest is 0. For example, for testing a correlation coefficient, the null hypothesis would be H_0: $\rho = 0$. We then compute a test statistic and corresponding *p*-value. The *p*-value indicates the probability of obtaining the observed data if the null hypothesis were true. If the *p*-value is very small (typically $p < .05$), we reject H_0, and conclude that there is a significant result. Conversely, if $p > .05$, we fail to reject the H_0, and the result is considered non-significant. It is most common to report *p*-values associated with a non-directional or two-tailed test, where significance does not depend on the direction of the result (i.e., either a positive or negative correlation would be significantly different from 0).

It is important to note that non-significance is a statement of uncertainty. If a result is not significant, the data do not provide enough evidence to conclude the results is different from 0. This is very different from concluding that the validity is 0. In other words, non-significance does not mean you should accept H_0. A common mistake in interpreting significance tests is to conclude that a predictor is unimportant because the test is non-significant. This tendency to treat a finding as true if $p < .05$ and false if $p > .05$ is not appropriate.

Statistical significant tests tend to be sensitive to sample size. For small samples, sampling errors tend to be large. Consequently, a particular result (say $r = .3$), might be due to *either* sampling error or a true non-zero correlation. With small samples, the data cannot tell the difference between these alternatives. Conversely, with very large sample sizes, even trivial results might be statistically significant. With a large enough sample size, a relationship of $r = .01$ might be statistically different from 0, but that does not make it practically important or useful in prediction.

There has been much criticism of statistical significance testing in the scientific literature, largely driven by the complicated logic and common misinterpretation of *p*-values and significance tests. In 2016, the American Statistical Association published a formal statement to address several common misconceptions of significance tests (Wasserstein & Lazar, 2016). They offer several principles to guide the interpretation of *p*-values:

1. *P*-values provide a measure of how incompatible the data are with the null hypothesis. The smaller the *p*-value, the more statistically incompatible the data are with the null hypothesis. If the underlying assumptions used to calculate the *p*-value hold, this incompatibility can be interpreted as casting doubt on or providing evidence against the null hypothesis.
2. *P*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

3. No decisions, whether scientific conclusions or business decisions, should be based only on whether a *p*-value passes a specific threshold. Interpreting relationships as true or false based on whether they fall below or above a "bright-line" rule such as $p < .05$ can lead to poor decisions. "The widespread use of 'statistical significance' (generally interpreted as '$p \leq .05$') as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process." (Wasserstein & Lazar, 2016)
4. Proper inference requires full reporting and transparency. Conducting multiple tests and then only reporting those analyses that were significant (known as cherry-picking or '*p*-hacking'), compromises the valid interpretation of those results. At minimum, valid scientific conclusions require knowing how many analyses were conducted and how those analyses were selected for reporting.
5. A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result. A trivial effect might be statistically significant if sample size is large enough. Conversely, large effects might produce non-significant findings when sample size is small.
6. By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis. A large *p*-value indicates the data are consistent with the null hypothesis, but there may be other models that are also compatible with the data.

Flagging Significant Results

When tabling results, it is common to use superscripts to indicate significant findings, often with different symbols depicting differing levels of significance. A benefit of this approach is that it avoids clutter and enhances the readability of slides, especially when many results are reported. Common superscripts are: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

When space permits, it is generally preferable to report the exact *p*-value rather than the ranges reflected in the significance flags (Aguinis et al., 2010). Reporting exact *p*-values conveys more information about the degree to which that data are inconsistent with the null hypothesis and allows the reader to apply a decision rule that differs from that of the researcher. Additionally, it makes more transparent the arbitrary distinctions near cutoff values (e.g., .052 vs. .049).

Significance flags can be an efficient mechanism for concisely communicating statistical significance in briefing presentations where additional statistical information might clutter the display and distract from the central message. In full technical reports, it is recommended that exact *p*-values and confidence intervals be reported.

1.4 Relevant Statistical Results for Criterion-Related Validity

When reporting validity results for individual assessment tools, at minimum, the following information should be included:

- Validity coefficient
- Significance level and confidence interval
- Information on adjustments for statistical artifacts

Each of these is discussed in detail below.

1.4.1. Validity Coefficient

The most common way to represent criterion-related validity evidence is through the correlation coefficient. This statistic provides a standardized index of the strength of linear relationships between two variables (predictor and criterion). That is, whether individuals with high (low) scores on the predictor tend to have higher (lower) scores on the measure of job performance.

The correlation coefficient has a theoretical range from -1 to 1 with 0 indicating a lack of relationship, 1 indicating a perfect positive relationship and -1 a perfect negative relationship. Figure 1 depicts data with varying degrees of correlation typical of validation research.

In some contexts, negative correlations are expected and the evidence provided can be just as strong as positive correlations. For example, when predicting counterproductive work behavior, we would want strong negative correlations between conscientiousness and the number of disciplinary actions. In other cases, a mix of positive and negative correlations might be expected. For example, individuals who experience low levels of Stress Under Pressure (SUP) would be expected to have higher performance than those who experience high levels of SUP. Thus, it is important to report the direction of the relationships and to highlight findings that are in a direction inconsistent with the theoretical rationale for the measure.

In order to make interpretation more straightforward, it can be helpful to reverse score measures that reflect negative characteristics. That way all correlations are expected to be in the same direction, and negative correlations would reflection relationships inconsistent with theory. If predictors are reverse coded, it is important that this be clearly noted in research reports. In order to reverse code a variable while maintaining the original range of scores, the following conversation can be used,

$$Score(Reversed) = Min + Max - Score \qquad (1)$$

where *Score* is the person's score on the original predictor variable, and *Min* and *Max* are the minimum and maximum scores that can be obtained on the predictor measure.

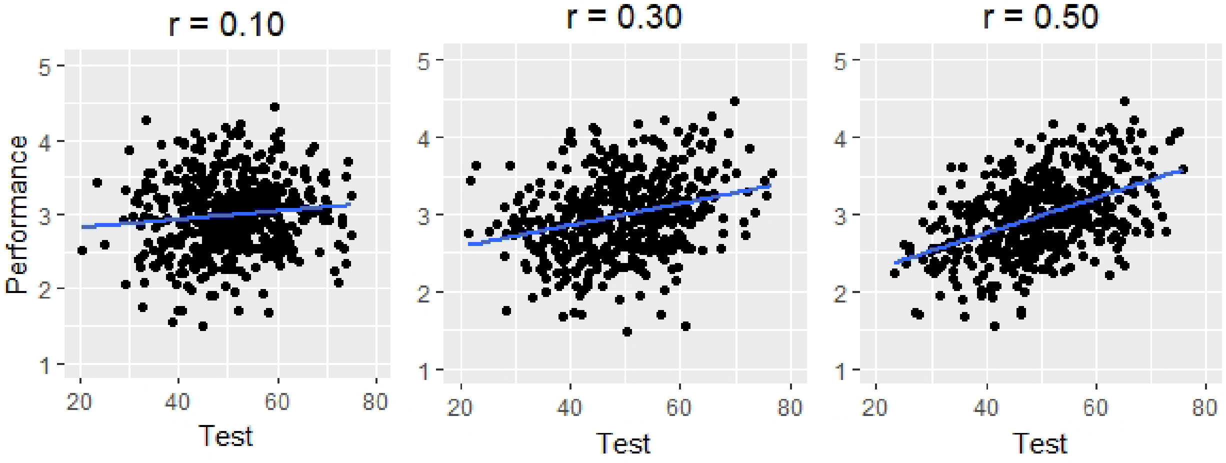


Figure 1. Scatter plots of data with varying levels of correlation

Correlation Benchmarks

Although the correlation coefficient has a standardized scale, it does not have a particularly intuitive scale. Those who are unfamiliar or do not regularly conduct statistical analyses may have difficulty determining whether a particular value, say $r = .2$, indicates a useful level of validity. Additionally, individuals unfamiliar with typical levels of validity might have unrealistic expectations. For these reasons, it is useful to have professional benchmarks that serve as a guide to interpreting the usefulness of results.

The U. S. Department of Labor (1999) provided a guide for interpreting validity coefficients for individual predictor variables (see Table 1). These values are intended to serve as general guidelines, and whether a particular value is acceptable depends on the context. The Department of Labor (1999) notes that the validity coefficient should be considered alongside several other factors, including: the level of adverse impact, the number of applicants relative to the number of openings, the cost of a hiring error, the cost of the selection tool, and the probability of hiring qualified applicant based on chance alone.

Table 1. Department of Labor Guidelines for Interpreting Validity Coefficients of Individual Tests

Validity Range	Interpretation	% Variance
Above .35	"Very beneficial"	> 12%
.21 - .35	"Likely to be useful"	4%-12%
.11 - .20	"Depends on circumstances"	1%-4%
Below .11	"Unlikely to be useful"	< 1%

Coefficient of Determination

Another way to convey the magnitude of correlation is through the coefficient of determination (r^2). The square of the correlation conveys the proportion of variance of the criterion measure that is explained by the predictor. For example, a validity of $r = .32$ indicates that the predictor explains 10% of the variance in job performance. Typical values for the proportion of variance explained are summarized in Table 1.

Statistical Significance and Confidence Intervals

A *t*-test can be used to evaluate the null hypothesis of no relationship (H_0: $\rho = 0$). The necessary statistics can be obtained from most statistical packages. For example, to obtain the correlation between variables *x* and *y* using the *cor.test()* function from the *stats* package in *R*, the command

```
cor.text(x,y)
```

would yield the following output,

```
        Pearson's product-moment correlation
data: x and y
t = 2.9166, df = 98, p-value = 0.004387
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.09126199 0.45383936
sample estimates:
     cor
0.2826141
```

When reporting results in text, include the value of the *t*-statistic, degrees of freedom ($df = N\text{-}2$) and the *p*-value. For example,

$$r = .28\ [.09,.45],\ t(98) = 2.92,\ p = .004, \tag{2}$$

indicates a validity coefficient of .28 with a 95% confidence interval ranging from .09 to .45, the value of the *t*-statistic with 98 *df* is 2.92, and the *p*-value is .004, which would be considered statistically significant at $\alpha = .05$.

In tabled results and presentations, a condensed presentation is used, reporting only the correlation coefficient, confidence interval or standard error, and one or more asterisks indicating the level of significance. A footnote on the table should indicate the sample size and provide a key to interpreting asterisks. The two formats are presented below (only one of these should be included).

Table 2. Brief validity presentation formats

Predictor	r [95% CI]	r (SE)
Test A	.28 [.09,.45]**	.28 (.10)**
Test B	.22 [.03,.40]*	.22 (.10)*

$\underline{N}$ = 100, $^{*}p$ <.05, $^{**}p$<.01

Statistical Adjustments

It is common for validity coefficients to be corrected for attenuation due to statistical artifacts, such as measurement error or range restriction. These adjustments are discussed in greater depth in Ployhart (2020). While these adjustments are common in validation research, many researchers are skeptical of the ability of post-hoc statistical corrections to overcome the weaknesses of poorly conducted research (e.g., LeBreton, Scherer & James, 2014; Schmitt, 2007). Thus, care should be taken that adjusted results are not viewed as attempting to overstate the evidence for validity. This can be achieved primary through transparency regarding the adjustments applied and their impact on results.

The corrections applied should be consistent with the available data and the purpose of the research. It is typically appropriate to adjust for measurement error in the criterion but not the predictor, because the goal of the research is to demonstrate the operational validity of the predictor measure and measurement error is an inextricable part of that measure. Similarly, adjustment for range restriction should apply a correction formula that matches the nature of the process that generated the restricted range of scores (Beatty, Barratt, Berry, & Sackett, 2014; Sackett & Yang, 2000).

It is important to be fully transparent regarding the statistical adjustments that are used. The technical report should identify which corrections were applied, the specific values used in the correction, and how those values were obtained. In the absence of data on artifacts in the current context, corrections are sometimes made using typical values obtained from the selection literature. In such cases, it should be noted that corrected validities are speculative. In general, both corrected and uncorrected validity should be presented in technical reports. For brief formats or with non-technical audiences, reporting only the corrected result is acceptable, as long as the nature of the correction is clearly noted.

When applying statistical corrections, it is important to account for these corrections in significance tests and confidence intervals. Using a corrected correlation in standard formulas for statistical significance testing can result in inaccurate conclusions. Corrections for statistical artifacts increase the standard error of the corrected correlation (Hunter & Schmidt, 2004). The higher validity produced by the corrections tends to be offset by the increased standard error, such that conclusions about statistical significance are unchanged. The simplest way to address this issue is to conduct significance tests on the uncorrected validity coefficient.

The increased uncertainty created by statistical adjustments must also be addressed when computing confidence intervals. The most straightforward approach is to compute a confidence interval on the uncorrected correlation, and then apply the statistical adjustment to the endpoints of this interval (Hunter & Schmidt, 2004). Although this approach requires some hand calculation, it can be readily adapted to whatever statistical correction is applied.

Using the previous example, we have an uncorrected correlation of .28 with an 95% confidence interval of [.09,.45]. Say we have determined that the reliability of our performance measure is r_{yy} = .64. The correction for unreliability can be applied to the estimate as well as the endpoint so of the confidence interval, as illustrated in Table 3.

Table 3. Calculation of confidence intervals on a corrected validity coefficient.

	Lower CI Limit	Estimate	Upper CI Limit
Uncorrected	.09	.28	.45
Corrected	$\frac{.09}{\sqrt{.64}} = .11$	$\frac{.28}{\sqrt{.64}} = .35$	$\frac{.45}{\sqrt{.64}} = .56$

Thus, the results for the corrected correlation and its 95% confidence interval should be reported as $r = .35$ [.11, .56].

1.4.2. Multiple Regression Analysis

In many cases, validation studies examine a collection of tests that are intended to be used together as a test battery. In such cases, multiple regression analysis can provide useful insights into how the collection of predictors function together. Multiple regression analysis involves creating an optimally-weighted composite of predictor variables. That is, each predictor is assigned a weight (the regression coefficient) and the weighted sum of the predictors can be used as summary variable, combining information from all of the predictors into a single score.

The results of a multiple regression analysis can be used to address two types of research questions:

- Overall predictive validity using the set of predictors, and
- Incremental validity for each predictor

Validity for the Set of Predictors

Two related statistics can be used as an overall index of how well a set of tests predict the criterion variable. The coefficient of determination (R^2) indicates the proportion of variance in the criterion variable accounted for by the set of predictors. The square root of this value is the multiple correlation coefficient (Multiple R), which represents the correlation of an optimally-weighted composite of the predictors with the criterion variable. The Multiple R is directly comparable to a validity coefficient, and therefore is the preferred method of presenting the overall predictive power of a set of tests.

If a set of predictors were all uncorrelated with each other, the R^2 for the set of predictors would be equal to the sum of the squared validity of each predictor. However, because predictors are usually somewhat intercorrelated, there is some redundancy in their prediction and the overall variance explained is typically less than the sum of the individual r^2 values. For this reason, it is important to assess how well the set of tests predict as a group, as well as individually.

A statistical significance test can be conducted on the model as a whole, testing the null hypothesis that the population multiple correlation is 0, or equivalently that none of the predictors are related to the criterion variable. An overall model F-test is reported in the output of most statistical packages. It is also possible to report a confidence interval on the Multiple R^2 (Cohen, Cohen, Wes, & Aiken, 2003) although this is less likely to be included in standard packages.

While the Multiple R is a useful summary statistic, it will often overestimate the level of validity that will be obtained when a battery is used in practice. Steps should be taken to avoid overstating the strength of the validity evidence.

First, the predictor weights used in the validation should match how test scores will be used in practice. The Multiple R from the regression equation represents the validity of a composite computed using the regression weights (which are computed to optimize prediction). Often, an operational selection system will use some other weighting scheme, for example, giving all predictors equal weight, or applying rationally derived weights (Hattrup, 2012). In such cases, the Multiple *R* may overestimate the operational validity, although the difference in the validity estimate is often quite small (Bobko, Roth, & Buster, 2007). Nevertheless, it would be best to report validity based on a composite score using the operational weights.

A second concern regarding the Multiple *R* is that regression weights have a tendency to overfit the data. Predictors are weighted to optimize prediction within the dataset used for the analysis and may pick up on idiosyncrasies that improve prediction for this sample but will not generalize to other applicants. Consequently, the Multiple *R* obtained in the regression analysis will tend to overestimate the level of prediction that will be achieved when the system is used in practice, a phenomenon known as ‘shrinkage’.

Cross-validation provides a methodology to obtain a more realistic estimate of expected validity. In cross-validation, the regression weights are estimated in one sample, and then a composite score based on these weights is validated in a second ‘holdout’ sample (Schmitt & Ployhart, 1999). Because obtaining two large samples for the purpose of validation is often impractical, a number of alternative strategies exist for estimating the cross-validated Multiple *R*. Computationally intensive methods, such as K-fold cross-validation, involve conducting several replications of the analysis where cases are alternatively used for the estimation or holdout samples in different replications (Putka, Beatty, & Reeder, 2018).

It is also possible to numerically approximate the cross-validated results without actually conducting a cross-validation study, by computing a so-called ‘adjusted’ R^2 or Multiple *R* statistic. Two different types of adjusted R^2 are available (Raju, Bilgic, Edwards, & Fleer, 1997). Estimates of the population R^2 reflect the R^2 that would be expected if using the population regression equation. Estimates of population cross-validity represent the R^2 that would be obtained if the weights derived in the current sample were used to predict outcomes for a new set of individuals from the same population. A variety of such adjusted R^2 estimates exist (Raju, et al., 1997, 1999). While many software packages report an adjusted R^2, this is typically the Wherry/Ezekiel formula, which is an estimate of population R^2, whereas an estimate of population cross-validity is more relevant to validation research (Schmitt & Ployhart, 1999).

While estimates of population cross-validity are not as common in statistical software, they are fairly easy to calculate by hand. The Burket formula is an attractive option due to its simplicity and the fact that it directly estimates the cross-validated Multiple *R*. Many other estimates, including the Browne formula, estimate the cross-validated *squared* Multiple *R*. The Burket formula is,

$$Burket\ Adjusted\ R = \frac{(NR^2-k)}{R(N-k)} \qquad (3)$$

and the Browne formula is,

$$Browne\ Adjusted\ R^2 = \frac{(N-k-3)R^4+R^2}{(N-2k-2)R^2+k} \qquad (4)$$

where *R* is the multiple *R* obtained on the current sample, *N* is the sample size, and *k* is the number of predictor variables.

A key determinant of shrinkage is the number of predictors (*k*). It is common for researchers to start with a larger number of predictor variables and then remove some based on the results of the analysis. The proper *k* to use in adjustment formulas is not the number in the final model, but rather the total number of predictors considered in the analysis (Schmitt & Ployhart, 1999). This is true regardless of whether predictors are selected using a formal procedure (e.g., stepwise) or based on the researcher's judgment after inspecting the correlation matrix. Consequently, it is generally best to compute shrinkage-adjusted R^2 statistics by hand, where the correct number of predictors can be specified.

Incremental Validity

In addition to the overall validity of the set of predictors, it is useful to examine the unique contribution of each predictor. Multiple assessments are included in a battery because they are expected to account for unique aspects of performance. Therefore each predictor in a battery should add value to the overall score.

In practice, many predictors are at least somewhat correlated with each other. When predictors are correlated, part of their relationships with the criterion overlaps and this shared prediction decreases the unique contribution of the predictor. Consequently, if tests in a battery are highly correlated, a test that is highly predictive when considered alone might add little to the overall validity of the battery.

The unique contribution of a predictor can be indexed through regression coefficients or incremental validity statistics. Regression coefficients are the weights assigned to predictors in the regression equation. Raw regression coefficients indicate the slope of the prediction line for one predictor when all other predictors are held constant. Raw coefficients can be difficult to interpret because they are influenced by the scaling of the variables. Consequently, researchers often interpret standardized coefficients, which are simply the regression weights that would be obtained if all variables were standardized before running the analysis. The relationship between raw and standardized coefficients is a function of the standard deviation of the predictor (SD_X) and criterion variables (SD_Y). This relationship can be used to convert from one metric to the other,

$$std.b = raw\ b\left(\frac{SD_X}{SD_Y}\right) \qquad (5)$$

$$raw\ b = std.b\left(\frac{SD_Y}{SD_X}\right) \qquad (6)$$

Given their similarity to validity coefficients, many researchers prefer to interpret standardized coefficients. However, the relative usefulness of raw versus standardized coefficients has long been an issue of contention among statisticians (e.g., Pedhazur & Schmelkin, 1991). For example, if analyses were conducted in two separate samples, one with a larger range of predictor scores (larger SD_X), then the same relationship in terms of the raw regression equation could produce very different standardized coefficients. This sensitivity to differences in variability creates challenges when attempting to compare standardized coefficients across samples.

Statistical significance tests can be conducted on the unique contribution of each predictor, testing the null hypothesis that the population regression coefficient is zero. A *t*-test, standard error, and confidence interval on each raw regression coefficient are generally available in the output of statistical software. Confidence intervals on standardized coefficients can be approximated by applying the standardization formula, to the endpoint of the confidence interval for the raw coefficient (see Jones & Waller, 2013, for an alternative procedure).

A second way to assess the contribution of a variable is to calculate the change in overall validity when that variable is added to the model. This is achieved through a hierarchical regression analysis, which involves estimating a series of models, where one or more additional variables is added at each step. Say we have two predictors, Arithmetic Reasoning (AR) and Reading Comprehension (RC) and we want to determine the incremental validity of RC. We estimate two regression models: a reduced model with only the non-focal predictor (AR) and then a full model with both predictors.

Reduced Model: Y’ = b0 + b1 AR, Multiple R = .2
Full Model: Y’ = b0 + b1 AR + b2 RC, Multiple R = .3

The incremental validity for RC is the difference in the Multiple *R* for the two models, $\Delta R = .3 - .2 = .1$. We can also index incremental validity using R^2, $\Delta R^2 = .09 - .04 = .05$, which indicates that RC is able to predict an additional 5% of the variance in the criterion variable.

Incremental validity also can be assessed for sets of predictors. For example, say the current selection process is based on a set of cognitive assessments and you are considering the addition of a personality measure. In this case, it would be useful to compare the validity of the cognitive battery (Reduced Model) to a battery with both cognitive and personality scores (Full Model). As above, incremental validity would be computed from the difference in multiple R between these two models.

It should be noted that the order of entering predictors can impact the ΔR associated with a predictor and therefore the order of entry should be given careful consideration. In some situations, the goal of the research will suggest a logical ordering. If the purpose of the study is to evaluate the addition of new measures, these new measures would be added in Model 2. In most cases, however, there is not pre-specific ordering. In such cases, a separate set of analyses can be conducted for each predictor, entering all other predictors in Model 1 and adding the focal variable in Model 2. This would then be repeated with each predictor as the focal variable.

Reporting Regression Results

In technical reports, the text should describe all analyses leading to the final regression model:

- Describe preliminary analyses evaluating regression assumptions (e.g., normality, linearity, homoscedasticity or error variance).
- Specify how predictors were selected for entry into the final model and identify variables that were examined and later dropped from the analysis based on the results. If a data-

driven model building strategy (e.g., stepwise predictor selection) was used, describe the procedure and the full set of predictors considered.

Tables reporting regression results should include the following elements:

- Specify the criterion variable in the table title or in column headers if results for multiple outcomes are reported in the same table.
- Report the sample size in a table footnote.
- Provide easily interpreted descriptive labels for each predictor variable. Where space limitations require abbreviations, include definitions in the table footnote.
- Include all regression coefficients, including the intercept, control variables, and focal predictors. Regression coefficients are denoted by b, and the standard error by SE_b.
- For each predictor, report the raw regression coefficients, standard error, and 95% confidence interval.
- Report the results of the significance test on each predictor. Where space permits, report the t-statistic, df and p-value. Where space is limited, report only the p-value and note the df in a table footnote.
- Report either the standardized regression coefficient or incremental validity for each predictor. The standardized coefficient is frequently denoted by *Beta* or β. Incremental validity is denoted by ΔR.
- Include a summary of overall model statistics, including the Multiple R, standard error of estimate (s_e), and F-statistic, df and p-value for the overall model significance test.
- Include an estimate of cross-validated Multiple R, either the results of a cross-validation study or numerical approximation (adjusted R). If reporting adjusted R, specify which formula was used in the text or a table footnote.

Briefing reports and presentations will necessarily require reducing the amount of information presented. The report should indicate the full set of predictors examined and the process used to select the final model. A table should be included with the following information:

- Incremental validity or standardized regression coefficients, with confidence intervals.
- Use asterisks to indicate the results of the significance test on each predictor (e.g., *p <. .05, **p <. .05, ***p <. .001).
- Report the overall Multiple R and significance level on the overall model test, along with the cross-validated Multiple R.

Example

The following hypothetical dataset and analysis is provided to illustrate reporting practice related to multiple regression analysis. R code for generating all tables in this section is provided in Appendix B. The example involves a system for selecting individuals into a technical training program. Previously, the only requirement for admission was possession of a degree in a technical field. The validation study evaluated whether a battery of cognitive ability tests could identify individuals likely to succeed in training. Additionally, there was interest in determining whether adding a personality measure would further improve the screening process.

An initial set of predictors included 10 cognitive ability subtests of the Air Force Officer Qualifying Test (AFOQT) and 30 personality facets of the SDI. The tests were validated on a sample of 200 trainees. Based on an initial examination of correlations (not reported here), five predictors were selected for further analysis. The predictors included three cognitive tests: AR, RC and Block Counting (BC); and two personality scales: Self-Discipline (SDis) and SUP, reverse coded so that higher scores indicate better stress tolerance). The criterion variable is the grade in a training course, represented as a percentage of total possible points achieved. Descriptive statistics are presented in Table 4.

The validity of the set of predictors was examined using a multiple regression analysis predicting training performance from all five scores. The full regression results are reported in Table 5. For a briefing report, a condensed format in presented in Table 6

Table 4. Descriptive Statistics and Correlations Among Study Variables

	M	**SD**	**AR**	**RC**	**BC**	**SDis**	**SUP**	**Perf**
AR	13.71	3.96	1.00					
RC	17.22	3.96	0.65***	1.00				
BC	14.38	5.97	0.41***	0.29***	1.00			
SDis	59.41	6.22	0.10	0.12	0.24***	1.00		
SUP	48.97	8.22	0.22***	0.16*	0.29***	0.40***	1.00	
Perf	69.58	9.82	0.53***	0.43***	0.48***	0.34***	0.37***	1.00

N = 200, **p*<.05, ***p*<.01, ****p*<.001

Table 5. Prediction of training grade from arithmetic reasoning (AR), reading comprehension (RC), block counting (BC), self-discipline (SDis) and stress under pressure (SUP, reverse coded)

Predictor	**Coeff**	**SE**	**Beta**	**T**	**p**
(Intercept)	23.60	5.48		4.30	< .001
AR	0.74	0.19	0.30	3.98	< .001
RC	0.31	0.18	0.12	1.74	0.083
BC	0.39	0.10	0.24	3.89	< .001
SDis	0.28	0.09	0.18	2.94	0.004
SUP	0.17	0.07	0.15	2.40	0.017
R^2	0.44	0.05			
Multiple R	0.66				
Adj. R	0.45				
Residual SD	7.45				
F	30.36				
Df	5, 194				
P	< .001				

N=200. Adj R is the estimated population cross-validity (Burket, 1964).

Table 6. Concise summary of regression results

Predictor	Coeff	95% CI
Arithmetic Reasoning	0.3***	[0.15, 0.45]
Reading Comprehension	0.12	[-0.02, 0.26]
Block Counting	0.24***	[0.12, 0.36]
Self-Discipline	0.18***	[0.06, 0.29]
Stress Under Pressure (R)	0.15*	[0.03, 0.27]
Multiple R	0.66	[0.58, 0.73]
Adjusted R	0.45	

Note: Adjusted R is the estimated cross validity (Burket, 1964).

In addition to the overall validity, we are interested in the incremental validity due to different components of the selection system. First, we want to assess the validity of the cognitive tests relative to simply requiring a prior technology degree (Degree). Second, we want to evaluate the additional contribution of the personality facets. To address these questions, we estimate a sequence of three models, adding a distinct type of predictor at each step (Degree, cognitive, personality). The results are summarized in Table 7.

Table 7. Hierarchical regression analysis

	Coeff (SE)		
	Model 1	**Model 2**	**Model 3**
Intercept	65.65	47.86	26.35
Degree	7.86 (1.27)***	2.85 (1.33)*	1.50 (1.34)
AR		0.67 (0.20)***	0.69 (0.19)***
RC		0.22 (0.19)	0.24 (0.19)
BC		0.51 (0.10)***	0.39 (0.10)***
SDis			0.25 (0.10)*
SUP			0.17 (0.07)*
Residual SD	9.01	7.75	7.44
R^2	0.16	0.39	0.44
Multiple R	0.40	0.62	0.66
F	38.02	30.98	25.54
Df	1, 198	4, 195	6, 193
P	< .001	< .001	< .001
ΔR^2		0.23	0.05
ΔF		26.25	9.36
P		< .001	< .001

Coeff = unstandardized regression coefficients.

Advanced Regression Models
If models include higher-order effects, such as interactional or polynomial terms to represent curvilinear trends, additional care is needed to avoid misinterpretation of findings. When higher-order terms are included in a model, the coefficients on their components are easily misinterpreted. For example, if a model includes predictors X1, X2, and their product X1*X2, the coefficient on X1 is the conditional slope of X1 when X2 = 0. Because a score of zero may not exist in the data, and may not even be a possible score for many tests, this conditional slope is often uninterpretable. Additionally, the presence of a significant interaction indicates that the slope of X1 is not the same for all examinees. Therefore, it is beneficial to compute and report simple slopes at several levels of the moderating variable (e.g., at higher and lower levels of X2). Procedures for computing simple slopes are described in Cohen et al. (2003), and Preacher, Currran, and Bauer (2006).

A plot of simple slopes often facilitates interpretation of interaction effects. Simple slopes should be plotted at two or more meaningful levels of the moderator variable. Provide an explanation in the text of how levels of the moderator were determined, and apply labels that accurately reflect the interpretation of these levels. A common practice is to use 1 SD above and below the mean on the moderator variable, in which case the label should reflect the relative nature of these value (e.g., “lower” and “higher”). Absolute labels such as “low” and “high” should only be used if they can be linked to an established interpretation of those scores.

For polynomial or curvilinear models, plot the predicted outcome at multiple levels of the predictor representing the full range of the achievable predictor scale. Using five or more levels of the predictor is often needed to accurately convey the curvilinear trend.

Although statistical packages will compute standardized regression coefficients on interactions and polynomial terms, these values are not meaningful and should be excluded from tables.

Advanced statistical techniques, such as structural equation modeling, come with additional reporting requirements which are beyond the scope of the current report. A useful guide for reporting results from structural equation modeling can be found at https://apastyle.apa.org/jars/quant-table-7.pdf

1.4.3. Logistic Regression Analysis

Logistic regression is used to develop and evaluate prediction models when the criterion variable is dichotomous. One or more predictor variables are used to predict the likelihood that an event occurs, such as successful completion of a training program or receiving a commendation, or negative outcomes such as disciplinary actions or turnover.

In order to properly model a dichotomous outcome, logistic regression uses a non-linear transformation of the dependent variable. Specifically, the regression model predicts the *logit*, or the natural log of the odds of the event (see Hosmer, Lemeshow, & Sturdivant, 2013). The results of a logistic regression look much like that of linear regression. However, due to the transformation the relation between the predictors and the actual outcome is not directly obvious from the regression coefficients. Therefore, greater care is needed in interpreting logistic regression coefficients.

Overall Prediction

As in linear regression, we are interested in evaluating both the model as a whole as well as the contribution of individual predictors. The overall model can be evaluated for statistical significance using the likelihood ratio (LR) χ^2 test. When reporting the results of a logistic regression, it is customary to also report a statistic called the *deviance* (sometimes denoted -2LL), which is a measure of the degree of prediction error. While it is not interpreted directly, the deviance is used in the calculation of the LR test, and many other statistical indices.

Several different R^2-like statistics have been developed to serve as an analog to the coefficient of determination (Cohen et al., 2003). However, in logistic regression there is no simple index representing the proportion of variance accounted for by the model and there is no consensus on which measure is preferred. Two common measures are the Cox and Snell and the Negelkerke pseudo-R^2 statistics. It is important to note that the choice of statistic can make a substantial difference in the results. The Cox and Snell R^2 tends to be conservative, in the sense that its maximum value is often less than 1.0. The Negelkerke R^2 includes an adjustment for this conservatism, and therefore tends to produce larger values than the Cox and Snell R^2.

Information criteria such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are also commonly reported with logistic regression results. These statistics are not directly interpretable by themselves, and are only useful when choosing among alternative models. In such cases, the model with the lower AIC or BIC value is considered the better model. AIC and BIC are based on the deviance (which measures the degree of prediction error) and also includes a penalty for the complexity of the model. As such, these statistics tend to prefer more parsimonious models. If we add predictors (increasing complexity) without substantially decreasing the prediction error, the AIC and BIC will increase, suggesting that the smaller, more parsimonious model be selected.

Although less commonly reported in research using logistic regression, another useful measure of the model's predictive power is the classification accuracy. For example, if we build a model to predict successful completion of a training program, we can calculate the percent of individuals for which the model prediction was correct (i.e., those predicted to succeed who were actually successful and those predicted to fail who actually failed). This approach is similar to the methods for constructing expectancy tables, which will be described in more detail later in this report.

Predictor Contribution

The contribution of each assessment to the overall prediction can be determined from the regression coefficients. As in linear regression, the sign of the coefficient indicates the direction of relationship. Positive coefficients indicate that higher predictor scores correspond to higher probability of the event, while negative coefficients indicate decreasing probability with increasing predictor scores.

Regression coefficients can be evaluated for statistical significance using the Wald test, which can either be presented as a Z-test or a χ^2 test. Similarly, confidence intervals can be constructed using a normal distribution (i.e., the 95% CI is computed using $b \pm 1.96\ SE_b$).

Interpreting the strength of the relationship is more challenging in logistic than in linear regression. Due to the non-linear nature of the logistic regression model, regression coefficients do not have a simple, intuitive interpretation. Standardized coefficients are generally not reported in logistic regression.

Research using logistic regression sometimes converts regression coefficients into odds ratios in order to aid interpretation, although this will only be helpful for audiences that are familiar with odds ratios. The conversion uses the antilog (EXP) function,

$$OR = EXP(b) = e^b \qquad (?)$$

The EXP function is available in Excel and most statistical packages. The result is an odds ratio corresponding to a one-point increase in the predictor. Say the *AR* test has a regression coefficient of 0.24. Using Excel we calculate EXP(.24) = 1.27, indicating that for every 1 point increase in AR score, the odds of passing are 1.27 times higher. It is important to emphasize that this ratio refers to odds and does not represent the expected change in probability or proportion who pass. Because non-technical audience and many selection practitioners are unfamiliar with odds ratios, and may misinterpret odds as relative probability, we do not recommend reporting EXP(b).

A more easily interpretable way of presenting the strength of a variable's contribution is to plot the predicted probability of success for a range of levels of the predictor, while holding all other predictors constant at their means. A line graph showing the predicted probability as a function of the predictor can be a useful way to demonstrate the strength of relationship. We illustrate how to accomplish this in R; a similar process can be adapted for Excel or other software.

1. Select a focal predictor and the number of levels (k) of this predictor you wish to examine. In order to capture the non-linear trend, at least five levels should be selected, evenly spaced across the full range of the predictor variable. More levels will tend to produce a smoother looking graph.
2. Create a table with k rows and columns equal to the number of predictor variables. Fill in the selected level of the focal predictor. For all other predictors, fill in the mean of the predictor for all rows.
3. Using the coefficients from the logistic regression output, compute the predicted logit for each row.
4. Transform the logit into a predicted probability, using the formula,

$$P(X) = \frac{1}{1 + EXP\big(-LOGIT(X)\big)}$$

If using R, steps 3 and 4 can are simplified and combined by using the *predict*() function.

Example
Using the example described above, logistic regression could be used to predict whether individuals successfully complete the training program, using the same set of three cognitive and two personality scores. A complete summary of results is presented in Table 8. The results show that the set of predictors accounts for a substantial portion of the variance in training completion and three of the predictors: Arithmetic Reasoning, Block Counting, and Stress Under Pressure (reverse coded); were significantly and positively related to rate of completion.

A graph of the relationship for AR is provided in Figure-2. Alternatively, the relationships can be depicted via a bar chart showing the predicted probability of success for specific levels of the predictor.

Table 8. Results of logistic regression predicting completion of training.

	Estimate	SE	Z-test	P
(Intercept)	-7.58	2.5	-3.04	0.002
AR	0.24	0.09	2.84	0.005
RC	0.00	0.07	0.02	0.988
BC	0.14	0.04	3.12	0.002
SDis	-0.06	0.04	-1.53	0.125
SUP (R)	0.08	0.03	2.56	0.011
Negelkerke R^2	0.38			
AIC	148.37			
LR χ^2	52.18			
df	5			
p	< .001			

N=200

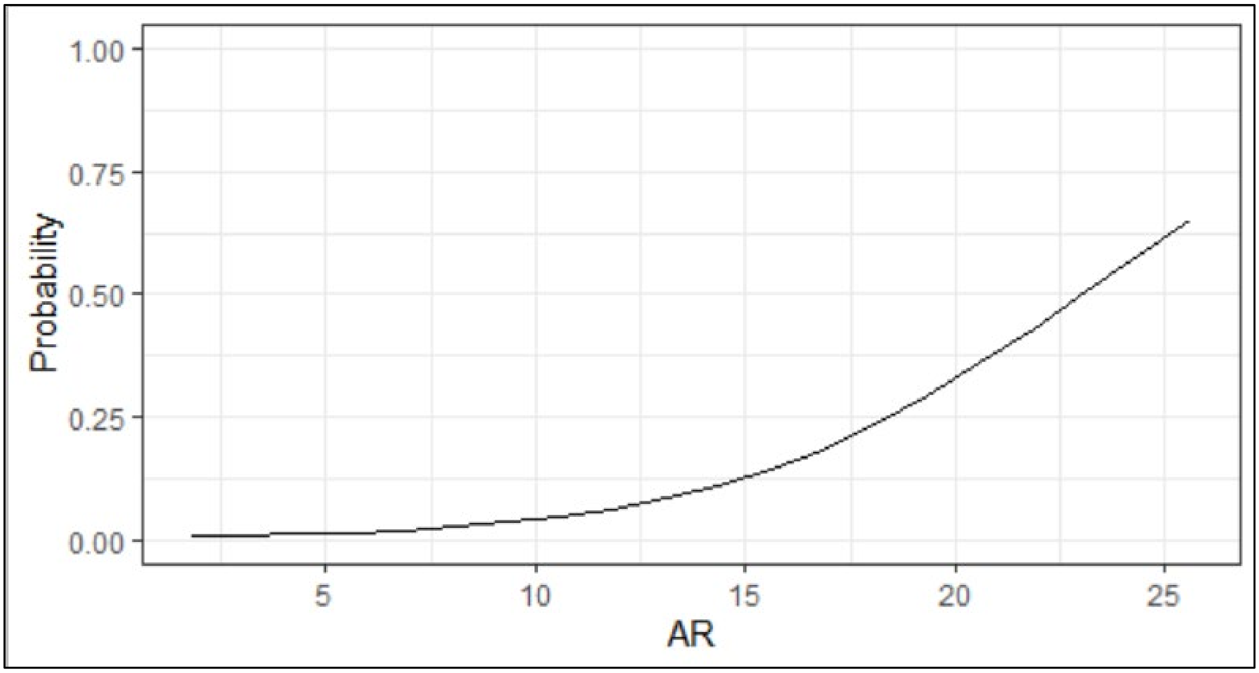


Figure 2. Relationships between Arithmetic Reasoning score and probability of completing training program

1.5 Group Differences

In addition to validity, it is also important to understand the impact of assessments on different demographic groups, particularly for historically underrepresented groups (Ployhart & Holtz, 2008). Data on group differences often plays a key role in discrimination claims and can be an important outcome for evaluating the effectiveness of diversity efforts. Therefore, validation reports should provide findings regarding group differences, where feasible.

It is useful to distinguish between two ways of operationalizing group differences on assessments: mean differences and passing rates. Mean differences between groups on an assessment reflect *potential* disparities. That is, if Group A tends to score higher on a test than Group B, then selection decisions made using that test will tend to favor Group A over Group B. However, the size of the disparity in test scores does not translate directly into the size of the disparity in selection decisions (i.e., the *achieved* disparity). Selection decisions tend to reflect a complex process involving many factors, including decisions about how to combine multiple tests, where to set cutoffs, the use of bands or score adjustments, etc., that are partly informed by the observed mean differences (Arthur, Doverspike, Barret ,& Miguel, 2013). When developing a new assessment procedure, the focus will be on potential disparities (mean differences), whereas when evaluating an operational assessment process, the achieved disparity will be more relevant.

Analysis of employment disparities involves a number of decisions regarding what individuals and demographic subgroups to include in analyses (Cohen, Fortney, & Tison, 2017). Data on demographic group memberships is often incomplete and inconsistent. Individuals are excluded from the analysis for a variety of reasons (incomplete data, not meeting minimum qualifications, etc.). These factors make it particularly important to maintain transparency regarding data cleaning and handling of missing data.

The list of relevant subgroups to examine is constantly evolving in step with societal norms and priorities. At minimum, subgroup analysis should be conducted by sex and for the racial/ethnic groups currently identified in federal equal employment opportunity reporting requirements. Race categories include: White or Caucasian, Black or African American, American Indian or Alaska Native, Asian, Native Hawaiian or Other Pacific Islander. Ethnicity is a separate classification for whether or not the individual is Hispanic/ Latino, regardless of racial category. It is generally best to report results for race and ethnicity comparisons separately, however, it is not uncommon to see race/ethnicity combined into a single variable, where all Hispanic/Latino are combined regardless of race and other racial categories only include the non-Hispanic/Latino individuals. An increasing number of individuals indicate they identify with multiple racial categories.

Comparisons typically are conducted among specific gender, race, or ethnic groups (Cohen et al., 2017). Aggregating multiple subgroups into an overall 'minority' group is not ideal, because the factors that lead to differential outcomes may not impact all subgroups in the same way. Typically, one group is identified as the reference group to which all other groups are compared. It is useful for the reference groups to be the majority group or the group that has historically been favored. Most often, men and Whites are used as the reference group; however, the Uniform Guidelines recommend using the groups with the highest selection rate.

1.5.1. Mean Differences

Mean difference can be presented in either raw score or standardized form. Because the range of scores varies across test and the scores often do not have an intuitive or familiar metric, differences in the raw score metric are often difficult to interpret. Therefore, the use of standardized mean differences is common. It should be noted that the concerns raised above regarding the use of standardized regression coefficients apply to standardized mean differences as well.

Cohen's *d* is a common standardized measure of the mean difference between two groups,

$$d = \frac{M_F - M_R}{SD_p}, \qquad (?)$$

where M_F and M_R are the mean scores for the focal and reference groups respectively, and SD_p is the pooled within-group standard deviation,

$$SD_p = \sqrt{\frac{(N_F-1)SD_F^2 + (N_R-1)SD_R^2}{N_F + N_R - 2}}. \qquad (?)$$

When comparing multiple focal groups to a common reference group (e.g., race comparisons), a single pooled SD_p should be computed using all subgroups. For *k* groups,

$$SD_p = \sqrt{\frac{\sum_k [(N_k-1)SD_k^2]}{(\sum_k N_k) - k}}. \qquad (?)$$

Cohen (1988) provided widely used benchmarks for small (.2), medium (.5), and large (.8) group differences. In the context of employment testing, large differences between demographic groups are not uncommon. It will be most useful to interpret the magnitude of *d* in the context of what is known about group differences for different types of selection procedures (Ployhart & Holtz, 2008).

Standardizing using the pooled SD rests on the assumption that the variability of scores is the same in both groups. In such cases, pooling data from both groups provides a more precise estimate. However, if there are known to be substantial differences in SD between groups, this can create inconstant standardization across comparisons. In such cases, it is better to use the SD from the reference group rather than pooling. Alternatively, standardization could also be performed using the SD obtained from test norms or other large representative samples.

Adjustment of standardized mean differences for statistical artifacts is possible, but less common than for validity coefficients. Because we are interested in the observed differences resulting from operational use of tests, correction for measurement error in the tests is usually not appropriate when examining group differences in test scores. If the sample used to estimate *d* differs from the candidate pool on which the test will be used (e.g., if *d* is estimated using incumbents rather than applicants), then correction for range restriction would be appropriate. However, in situations where the correction is needed, the information required for range restriction correction formulas is often unavailable. Additional information on correcting the standardized mean difference for statistical artifacts can be found in Hunter and Schmidt (2004) or Bobko, Roth, and Bobko (2001).

When reporting results, mean differences should be accompanied with some indication of statistical precision: either a standard error, a confidence interval, or the results of a statistical significance test. Significance testing can be conducted using the *t*-test for independent groups. While software used for the *t*-test will typically report the standard error and confidence interval for raw means, these cannot be used when reporting the standardized mean difference. The standard error for Cohen's *d* is easily computed by hand,

$$SE(d) = \sqrt{\frac{N_F+N_R}{N_F N_R} + \frac{d^2}{2(N_F+N_R)}} \qquad (?)$$

The standard error can be used to construct a confidence interval using $d \pm t_{\text{crit}} * SE$, where t_{crit} is the critical value of a *t*-distribution with $df = N_F + N_R - 2$, at the desired confidence level (e.g., $\alpha = .05$ for a 95% CI).

1.5.2. Passing Rate Differences

When analyzing the achieved disparities resulting from the operational use of a selection system, the analysis will focus on group differences in the passing or selection rate. A variety of statistics can be used to evaluate group differences in passing rates (Oswald, Dunleavy, & Shaw, 2017), the most common of which is the adverse impact ratio. Let N_i represent the number of applicants in group *i*, and NP_i the number of those individuals who receive passing scores on the assessment. The adverse impact ratio (AIR) is the ratio of the passing rates for the focal and reference groups,

$$AIR = \frac{{NP_F}/{N_F}}{{NP_R}/{N_R}} \qquad (?)$$

The adverse impact ratio is often associated with the four-fifths rule, but they are not the same thing. The four-fifths rule was suggested in the Uniform Guidelines as a guide to determine when a disparity was sufficiently large to merit scrutiny. Specifically, a violation of the four-fifths rule occurs when AIR < 0.8. The AIR is an effect size statistic that quantifies the degree of disparity on a continuum, whereas the four-fifths rule is a dichotomous decision rule applied to the AIR. Research reports should include the actual AIR, not just whether system violated the four-fifths rule.

There has been considerable criticism of the four-fifths rule in the personnel selection literature (Roth, Bobko, & Switzer, 2006), and contemporary adverse impact analysis tends to focus more on statistical significance than the four-fifths rule (Tonowski, 2017). The AIR is nevertheless a useful measure of the size in selection disparities.

The difference in passing rates can be tested for statistical significance using a *Z*-test for two independent proportions, or equivalently the chi-square test for independence between groups' membership and the pass/fail decision. When sample size is small, the Fisher Exact Test is often used. For more information on significance testing of group differences in passing rates, see Morris (2017).

Alternatively, we can construct a confidence interval around the AIR (Morris & Lobsenz, 2000). This will often be more useful than significance tests. Because many predictors are known to show group differences, testing whether a disparity is different from zero is of limited utility. We are often more interested in the magnitude of the disparity relative to other selection procedures or prior selection systems. Confidence intervals are beneficial in this context because they provide an index of the degree of uncertainty in the estimate due to limited sample size, without losing sight of the actual value of the AIR.

Other than the four-fifths rule, there are no established benchmarks for interpreting whether a disparity is small, medium, or large. Additionally, the values obtained, being ratios of ratios, are rather abstract and non-intuitive. One way to translate passing rate disparities into more concrete numbers is through shortfall analysis. The shortfall represents the number of minority candidates who were negatively impacted by use of the selection procedure, relative to a system that had no disparity. The shortfall is computed as follows:

1. Compute the passing rate for the majority group. The selection ratio (SR) for the majority grouo equals the number passing (NP) for the majority group divided by the number in the majority group)N) or $SR_{maj} = NP_{maj}/N_{maj}$
2. Multiply the majority passing rate (SR_{maj}) by the number of minority candidates ($_{Nmin}$). This is the expected number of minority candidates selected (EP_{mim}) under a neutral system (i.e., a system where the passing rate is the same for both groups), $EP_{min} = SR_{maj} * N_{min}$
3. The shortfall is the difference between the expected and actual number of minority candidates selected, Shortfall = $EP_{min} - NP_{min}$

Another statistic that can be used to index employment disparities is proportion of the selected candidates who are from the minority group (i.e., NP_{min}/NP). Unlike the AIR, which takes into account the number of minority candidates, the minority representation among the selected reflects both the passing rates on the selection procedure and the composition of the applicant pool. As such, it is of limited usefulness for evaluating the selection procedure. At the same time, it provides a direct measure of whether the recruitment and selection system as a whole is likely to achieve minority hiring goals.

1.6 Communicating Validation Results

Effectively communicating technical information like test validity to a broad audience can be challenging. The presenter must balance multiple goals, including educating the audience on professional standards, presenting results in a fashion that is easily understood, and minimizing the risk that the audience will misinterpret findings. Whereas scientific norms tend to favor providing more information in the service of transparency and reproducibility, these practices often result in reporting formats that can be overwhelming to non-technical audiences. Additionally, standardized reporting standards of scientific publications generally do not include contextual information that is of greatest interest to decision-makers in applied settings (Aguinis et al., 2010).

It is well known that large amounts of information can lead to cognitive overload (Mayer & Moreno, 2003). Comprehension can be improved by limiting content to essential information, keeping slides simple, and providing signals to help the audience process information. Some general recommendations for preparing slide decks include:

1. Limit text on slides. Bullet points help to organize ideas, whereas complete sentences increase processing demands.
2. Minimize graphics and text on the same slide. Oral narration works better than explanatory text on slides.
3. Avoid repetition between oral narration and text on slides.
4. Avoid extraneous information on slides (e.g., clip art, animated transitions, etc.)
5. Provide visual signals to guide processing of information on slides (e.g., text color, bold or italicized font, boxes around key content). Position text near corresponding parts of graphics.

A similar call for simplicity in graphics is made by Kuncel and Rigdon (2013):

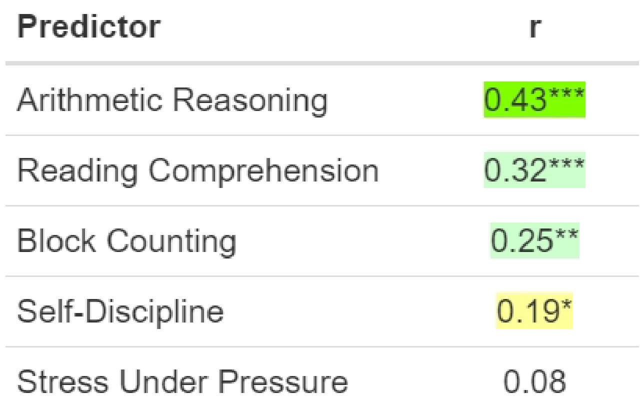
1. Keep graphs simple. Avoid decorative elements of graphs that do not convey additional information (e.g., 3D effects)
2. If possible, values should be labeled directly rather than through a legend.
3. Avoid multiple *y*-axes. In order to compare and contrast results for distinct outcomes, researchers sometimes plot distinct variables on the same graph (one scaled using the left *y*-axis, one scaled on the right *y*-axis. This practice should be avoided, because such graphs are cognitively demanding and prone to misinterpretation.

1.6.1. Presenting Tables

Tables provide a concise format to summarize large amounts of information and are quite useful in written research reports. However, in the context of a presentation, tabled information can easily become overwhelming. When including a table as part of a presentation, the presenter should carefully consider the intended message and include only information relevant to that message. This is a situation where the benefits of simple presentation need to be balanced against the professional responsibility to provide a full accounting of findings. This might be addressed by preparing supplemental slides or handouts that provide more detailed and complete information, while the presentation focuses only on key findings.

Presentation of tabled information can be facilitated by the addition of highlighting (e.g., bold, italics, color) that assists the audience in identifying patterns in the data. Color-coding validities based on the Department of Labor (DOL) levels of validity evidence can help quickly convey the degree of evidence. See Table 6-9 for an example.

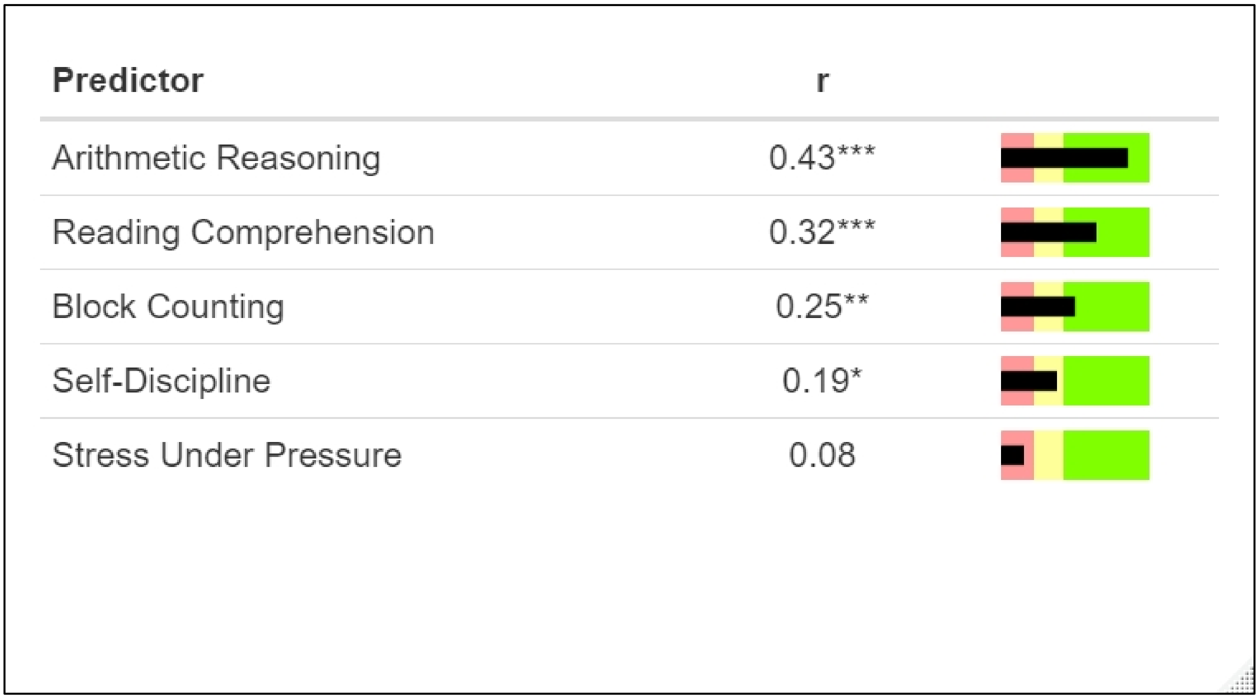
Table 9. Table of validities color coded by level of validity evidence



Predictor	r
Arithmetic Reasoning	0.43***
Reading Comprehension	0.32***
Block Counting	0.25**
Self-Discipline	0.19*
Stress Under Pressure	0.08

Similarly, graphical elements can be used to reinforce the numerical information. The bullet graph is one useful example. The bullet graph consists of a bar graph (representing the validity coefficient) plotted against a background with multiple colored regions representing distinct regions (DOL validity categories). An example is provided in Table 10.

Table 10. Table of predictive validities with bullet charts



Predictor	r	
Arithmetic Reasoning	0.43***	
Reading Comprehension	0.32***	
Block Counting	0.25**	
Self-Discipline	0.19*	
Stress Under Pressure	0.08	

DoD guidelines for interpreting validity: Red = "unlikely to be useful" (<.11); yellow = "depends on circumstances" (*r* = .11-.20), green = "likely to be useful" (*r* = .21-.35) or "very beneficial" (*r* > .35).

1.6.2. Expectancy Charts

A well-established method to communicate validity information is through the use of expectancy charts (Cucina, Berger, & Busciglio, 2017). An expectancy chart is a form of bar chart that depicts the level of a criterion variable for one or more ranges of scores on the predictor. For example, a chart might plot the percentage of high performers among candidates above vs. below the passing score on the test.

The expectancy chart seeks to simplify validity information in several ways. First, rather than referring to predictor level in terms of test scores (which will typically have little meaning to the audience), the expectancy chart presents categories representing meaningful score ranges with easily understood labels. For example, if a passing score has been established, it is common to present results for those who pass vs. fail the assessment. Alternatively, one might separate predictor scores into quartiles: bottom 25%, lower middle 25%, upper middle 25%, and top 25%.

Similarly, an effort is made to present the criterion in a metric that is meaningful to the audience. Because performance outcomes are often measured on a non-intuitive scale, it is useful to transform the criterion level into a percent of candidates expected to succeed, where success is defined as falling above a pre-defined cutoff on the criterion measure. For example, successful performance might be defined as achieving a rating that corresponds to meet expectations. Alternatively, we might compute the percentage who are high performance, defined as being among the top 20% on the criterion measure. The specific cutoff defining success will need to be determined in each selection context based on the goals of the assessment and the nature of the criterion variable.

By moving from continuous scores to classification decisions, we greatly simplify the interpretation of validity information. The concept of percentage successful is straightforward and directly relevant to decision-makers (Kuncel & Rigdon, 2013). Rather than reporting a validity using an abstract statistic like the validity coefficient (e.g., r = .35), we can report that among those who pass the test, 70% are successful performers, whereas among those who fail the test only 40% are successful.

Taylor-Russell Expectancy

Taylor and Russell (1939) provided a useful framework for understanding the relationship between the validity coefficient and probability of success. The model has three components: the validity coefficient, the passing rate on the test, and the base rate of success. Figure 3 depicts data with a validity coefficient of *(* = 0.5. The ellipse shows the 95% confidence region. The horizontal line represents the cutoff for successful performance; in this case, about 40% of all candidates are successful. The passing rate is reflected in the region of the ellipse to the right of the vertical line (regions B and D). The black vertical line represents a highly selective setting where the passing rate is about 15%. The red vertical line reflects a different scenario where the passing rate is around 50%.

The *expectancy* refers to the proportion of those who pass the test who also have successful performance (i.e., B/(B+D)). In other words, among those to the right of the vertical line, what proportion are in the upper right quadrant. With a strict cutoff (the vertical black line), only 15%

will pass the test and about 70% of those who pass are successful. A less stringent cutoff (the red vertical line) will pass 50% of candidates and only about 50% will be successful.

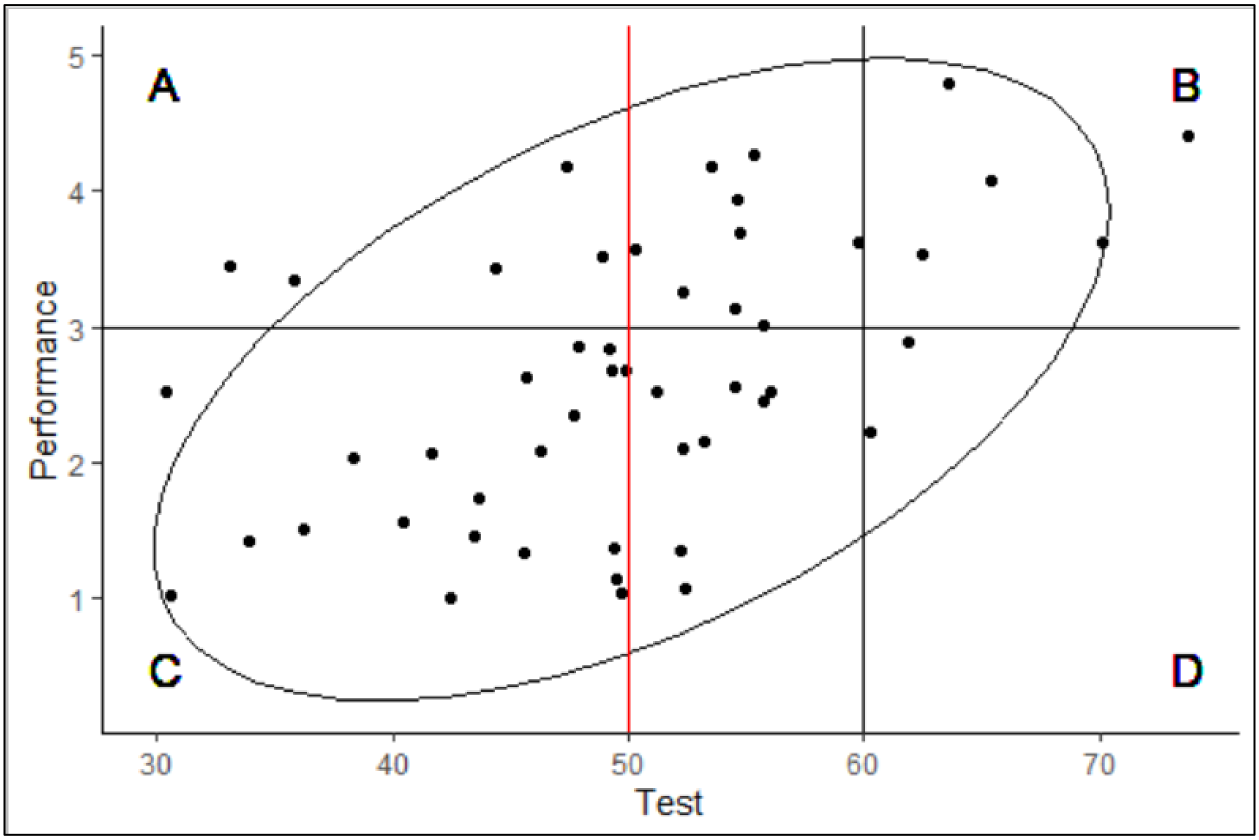


Figure 3. Quadrants of the test-performance relationship

Taylor and Russell (1939) showed how the expectancy can be calculated given the validity, base rate of success, and passing rate, and the density function of a bivariate normal distribution. They provided extensive tables that yield the probability of success for variance levels of the three parameters. More recently, Cucina et al. (2017) provided code for performing this calculation in R. A slightly modified version of this code is provided in Appendix C.

To run the Cucina et al. (2017) code, first copy the code in Appendix C to an R script file named "Cucina Expectancy Function.R", and save this file to the working directory. The code in this file creates a function *Expectancyfunc()*, which can be used to calculate the expected probability of success. Next, copy the following commands to a second R script file, changing the input parameters to reflect the situation, then run the code.

The code returns the proportion of examinees from a predictor range who are expected to fall within a criterion range. Both predictor and criterion ranges are defined by an upper and lower limit that must be specified. If no upper limit is desired, this value can be set to *Inf*. Similarly, -*Inf* can be used if no lower limit is desired.

```
# Setup
# The following line indicates the location of a code for the expectancyfunc() function.
source("Cucina Expectancy Function.R")

# Input parameters
validity <- .5                              # validity coefficient
passRate <- .15                             # proportion of examinees who pass the test
successRate <- .30                          # proportion of examinees with successful performance

# Intermediate calculations
predictorLowerCut <- qnorm(1-passRate)      # standardized test cut score
priterionLowerCut <- qnorm(1-successRate)   # standardized criterion cut score
predictorUpperCut <- Inf                    # upper limit of test score range (use Inf if no upper limit)
criterionUpperCut <- Inf                    # upper limit of criterion range (use Inf in not upper limit)

# Calculate expectancy and print result
expOut <- Expectancyfunc(r,predictorLowerCut,predictorUpperCut,
CriterionLowerCut,CriterionUpperCut)
cat("Expectancy = ", expOut$expectancy)  # print result
```

The function returns a list of results which is assigned to the object "expOut". The list contains 3 results. To reference a specific element, specify the list name, following by '$' and the element name. The three elements of are:

> *expOut$jtprob*: joint probability of an individual being selected and successful
>
> *expOut$xprob*: probability of an individual being selected (should be equal to passRate)
>
> *expOut$expectancy*: conditional probability of being successful if selected

Using the values listed above, the expectancy is 0.61. That is, among those selected (the top 15% of scores), 61% are expected to be successful (in the top 30% of performers).

The expectancy is typically presented in the form of a bar chart. Several examples are provided below. R code for creating an expectancy chart is provided in Appendix B. Additionally, an online utility for creating expectancy charts from data is described in Zhang (2018).

Confidence Intervals

Because the expectancy is based, in part, on the sample estimate of the validity coefficient, there will be uncertainty in the results due to sampling error in the validity coefficient. It is useful to convey this information in graphs via error bars (Kuncel & Rigdon, 2013). Cucina et al. (2017) recommend building a confidence interval around the expectancy by (1) finding the confidence limits of the validity coefficient and (2) calculating the expectancy separately for each confidence limit. Say the 95% confidence interval on the validity in the previous example is [.23,.67]. We add the following lines to the code above to calculate the confidence interval on the expectancy.

```
# Input confidence interval on validity coefficient
validityCI <- c(.23,.67)

# Expectancy using upper and lower bounds of confidence interval on validity
expL <- Expectancyfunc(validityCI[1],predictorLowerCut,predictorUpperCut,
                criterionLowerCut,criterionUpperCut)
expU <- Expectancyfunc(validityCI[2],predictorLowerCut,predictorUpperCut,
                criterionLowerCut,criterionUpperCut)

# print result
cat(paste(sprintf("Validity = %.2f, Pass Rate = %.2f, Success Rate = %.2f",
                validity, passRate, successRate),
sprintf("Expectancy [95%% CI]  = %.2f [%.2f, %.2f]",
                expOut$expectancy, expL$expectancy, expU$expectancy),
sep = "\n"))
```

Empirical vs. Distribution-Based Expectancies

When the validation data are available, an option would be to compute the probability of success directly from the data, rather than from the normal density function. That is, the researcher could simply calculate the number of examines in the data who pass the test and the proportion of those who are classified as successful based on their job performance.

Using the data in Figure 3, seven examinees passed the test using the cutoff score of 60 and five of these had performance in the successful range. This yields an expected success rate of 5/7 = 0.71, which is fairly close to the .61 obtained from the distribution-based analysis.

This empirical approach has the advantage of fewer assumptions. For the distribution-based approach, one must assume that the relationship between test scores and performance is linear and that the data follow a bivariate normal distribution. The empirical approach is particularly appealing when the outcome is categorical, given that the distribution-based method treats both variables as if they were continuous. For a dichotomous criterion (e.g., succeed/fail), imputing success rates from a bivariate normal distribution would be both unnecessary and less accurate.

However, Cucina et al. (2017) demonstrated that this empirical approach is more susceptible to sampling error, especially when sample sizes are small or the passing rate is low. Consider the stringent selection rule depicted in Figure 3. The empirical calculation of the expected probability of success would be based on only seven cases that were above the test cutoff. The resulting value would be very unstable and unlikely to replicate in a different sample.

Balancing these concerns, the distribution-based approach will be more useful in most situations. However, the empirical approach is likely to be useful when sample sizes are very large, when the outcome variable is categorical or when there is good reason to question the assumptions of linearity or bivariate normality.

Expectancy Comparison

A single expectancy result, considered in isolation, offers little help in understanding the usefulness of a selection procedure. Some form of comparison is needed. In the previous example, it was found that among those selected, 61% are expected to be successful. Whether this is considered a good or a bad result depends on the success rate that might be obtained through other means. Several variants of the expectancy chart are possible, differing in the specific comparison represented. Several possibilities are illustrated below and several additional options are described by Allred (1991).

One approach would be to compare the target test score range to other ranges. The simplest version compares the expectancy for those who pass the test to the expectancy for those who fail. Here we define two test score ranges, the first from negative infinity to the test score cutoff and a second from the test score cutoff to positive infinity. We can then use a bar chart to plot the expectancy for both ranges. To illustrate, Figure 4 presents expectancy charts for each of the validities reported in Table 4, using 60% as the success rate and a 30% pass rate on each test.

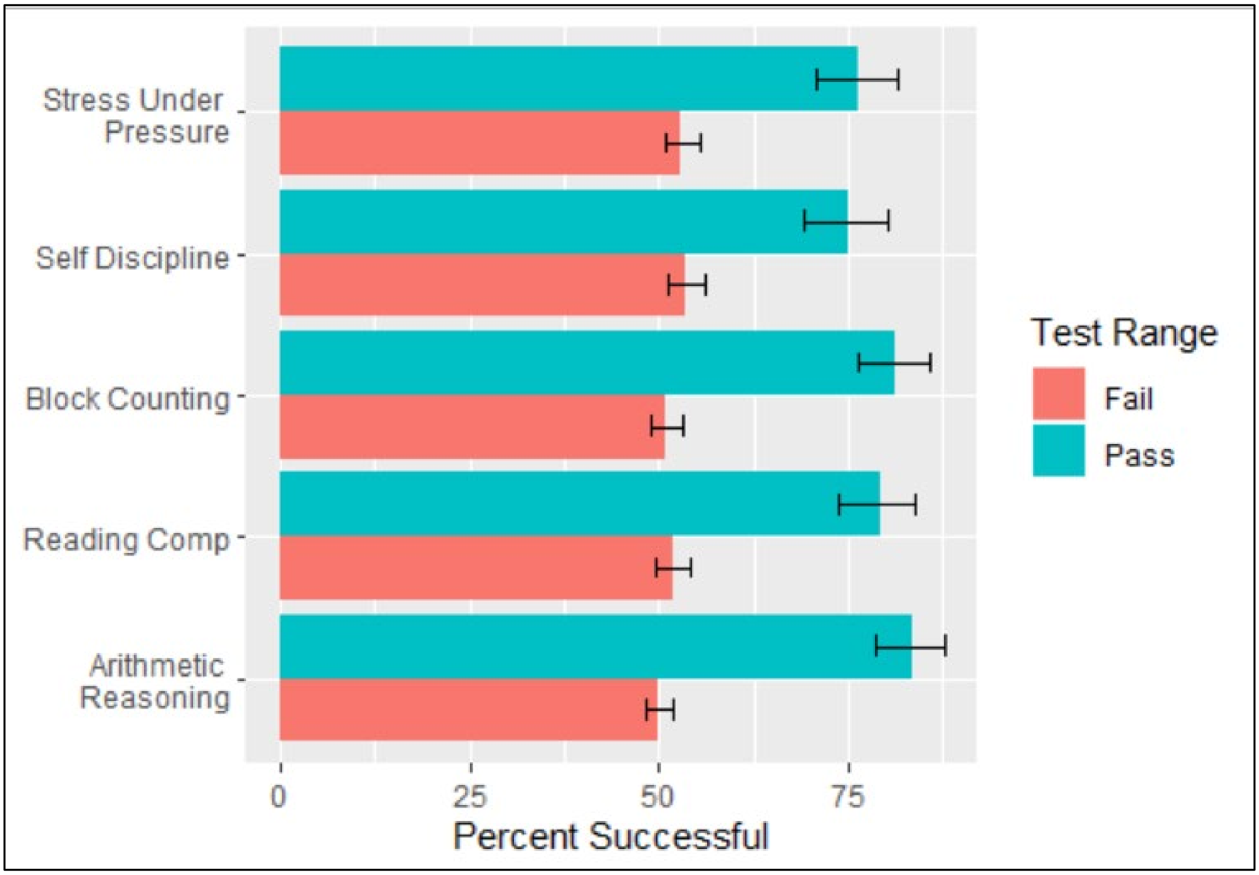


Figure 4. Expectancy Chart Showing Multiple Predictors

In addition to the pass/fail distinction, the expectancy chart might also show results for multiple ranges of scores for a single predictor. Because a primary goal of the expectancy chart is to simplify the presentation of validity information, the number of predictor ranges should be kept to the minimum needed in a particular context. For example, when setting a cutoff score, it can be useful to plot expected success for many test score ranges, because the goal is to pinpoint the optimal location of the cutoff score. However, in most contexts, such fine distinctions are unnecessary and multiple predictor score categories will only add to the cognitive demands of interpreting the graph. Figure 5 illustrates the expectancy charts with four predictor categories.

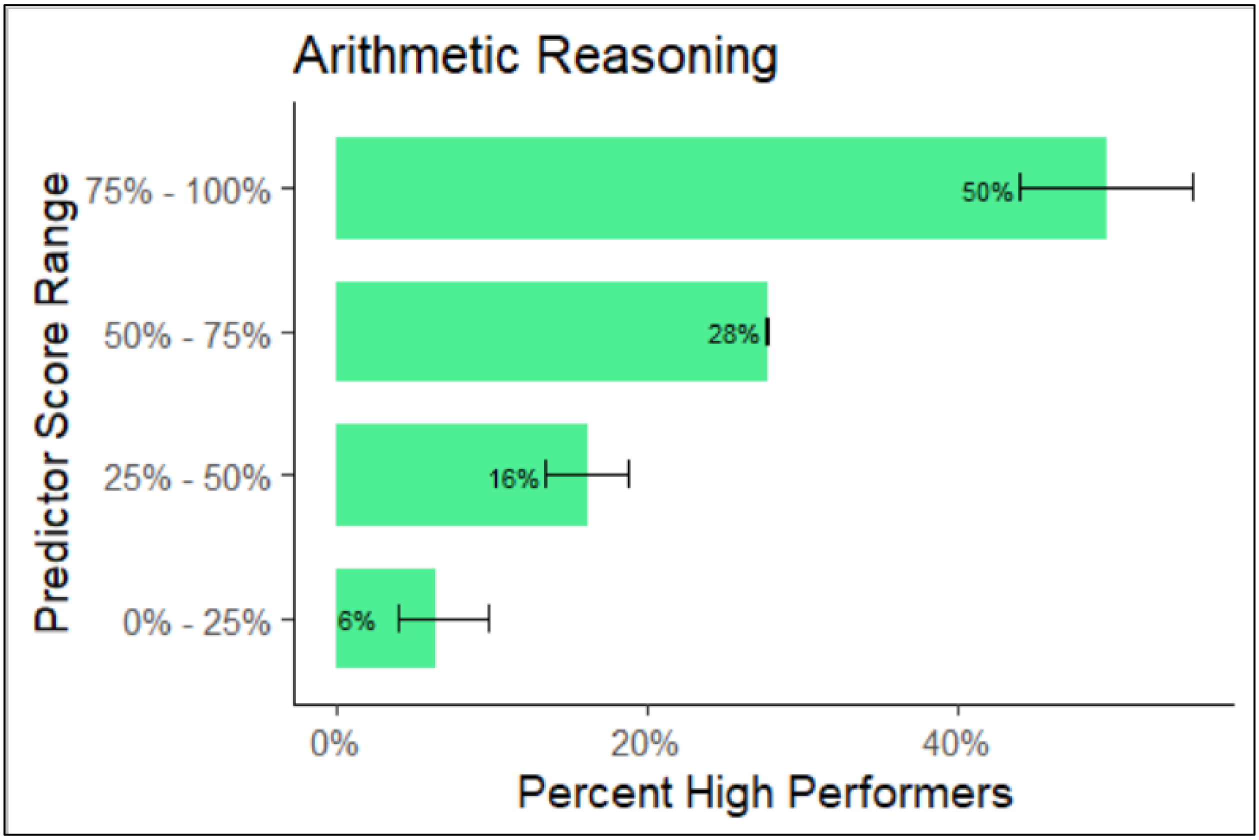


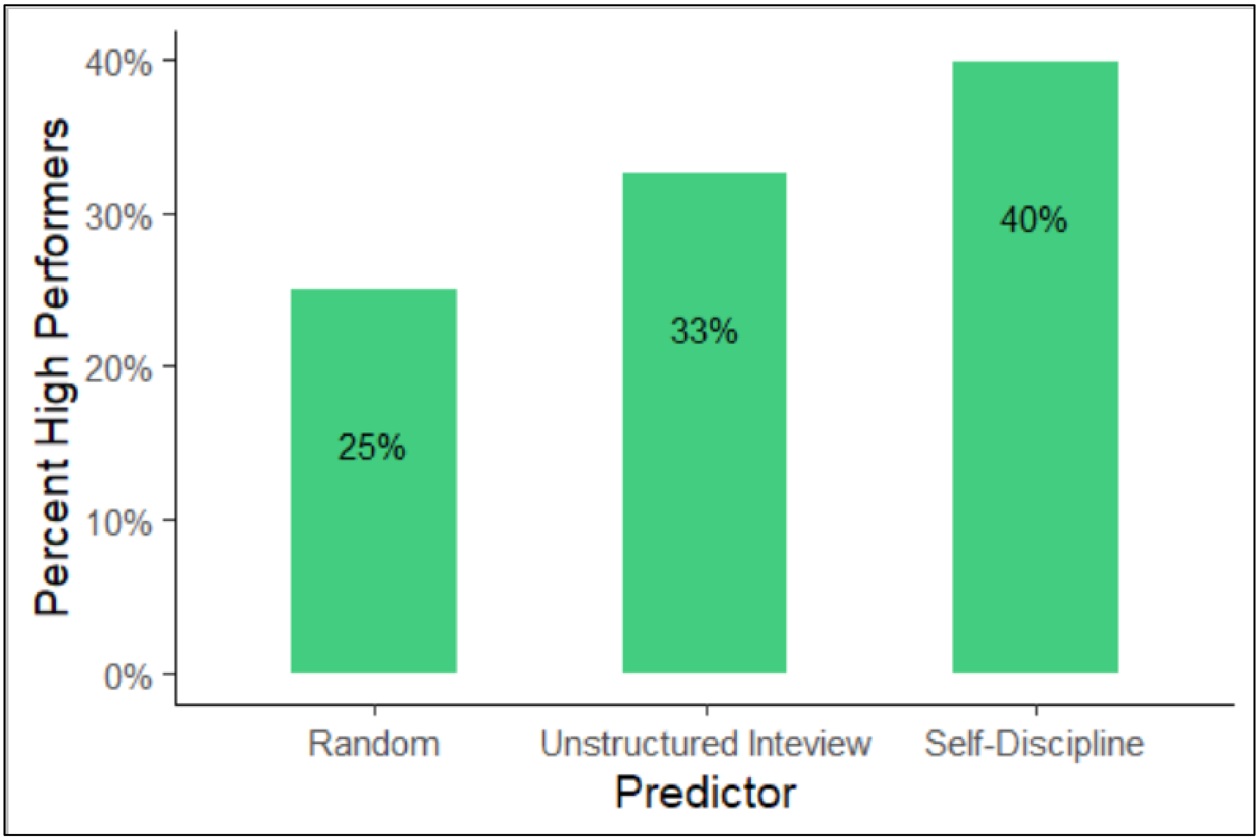
Figure 5. Expectancy charter for Arithmetic Reasoning four predictor levels
Note: Confidence intervals represent the expectancy associated with the 95% confidence limits on the validity coefficient, which produce similar expectancies near the predictor mean.

Another way to illustrate the benefit of a test is to compare its expectancy to that of other possible predictors. According to Zhang, Highhouse, Brooks, and Zhang (2018), graphic displays of validity information will be most informative when they show improvement in the level of validity. Improvement might be characterized relative to random selection, an existing selection system, or typical values for alternative predictors.

Typical values for alternative predictors can provide benchmarks by which to judge the performance of the test under consideration. Relevant values for other predictors might be obtained from prior validation research on similar jobs or from the general personnel selection literature (e.g., Schmidt & Hunter, 1998).

When using validities from the selection literature as a point of comparison, it is important the statistics are comparable in terms of statistical corrections. It would not be appropriate to compare a current uncorrected validity estimate to one adjusted for measurement error and range restriction. Care should be taken to identify reference estimates that involve the same statistical corrections that are used in the current study.

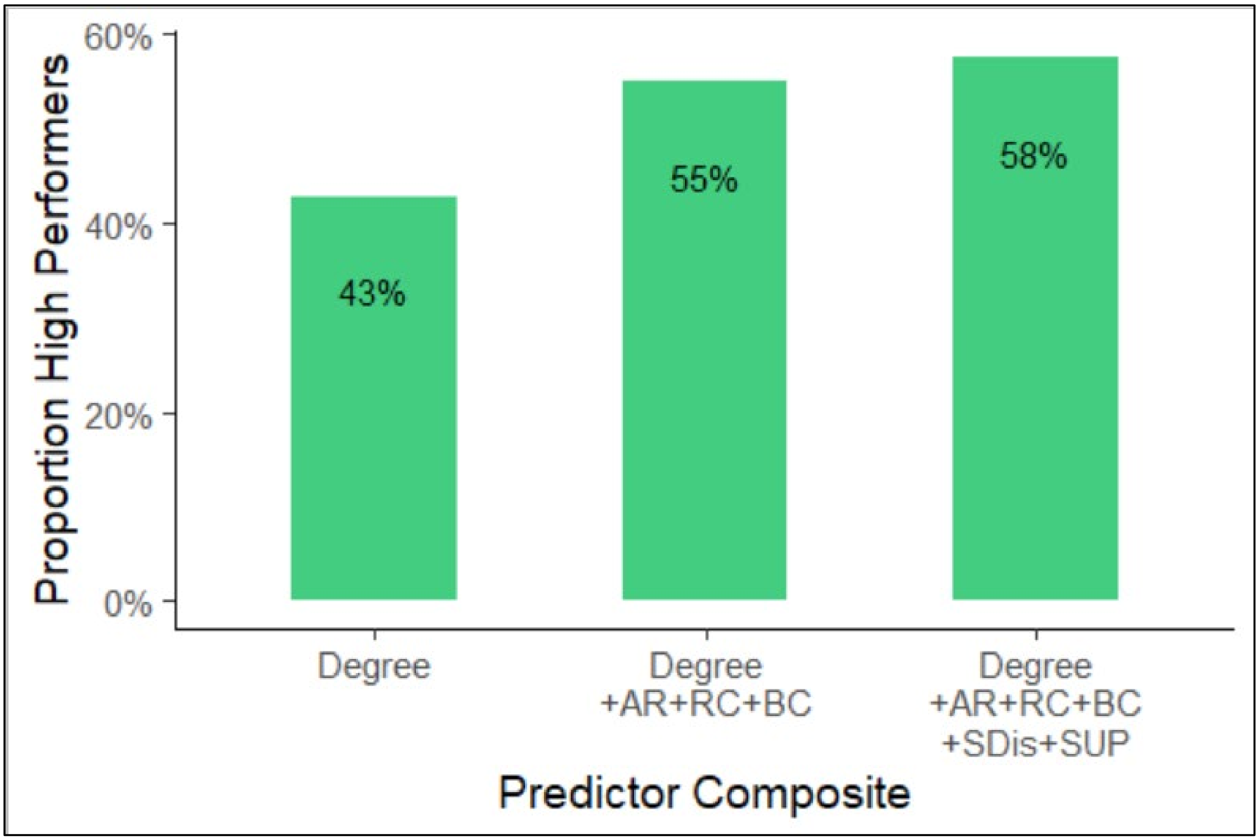
One useful point of comparison is the unstructured interview, which approximates what might be achieved through an informal selection process with no systematic testing. McDaniel et al. (1994) reported a mean uncorrected validity of .18 for unstructured interviews and validity of .33 after correction for criterion unreliability and range restriction. In the example described above, we obtained an uncorrected validity of .34 for Self-Discipline predicting training performance. Figure 6 displays the expectancy for random selection, unstructured interviews, and the Self-Discipline scale, with the top 25% of test scores selected and high performance defined as the top 25% of grades in the training program.



Note: Expectancy for the unstructured interview was based on a validity of .18 (McDaniel et al., 1994).

Figure 6. Expectancy of high performance when selecting candidates using SDI Self-Discipline Scale relative to no selection procedure (random selection) and an unstructured interview.

The comparison of alternate predictors also lends itself to questions of incremental validity in the context of assessing the unique contribution of predictors to a test battery. As described above in the section on regression analysis, a series of regression models is estimated, with one or more predictors added at each step. In the example above we examined the prediction of training performing of using prior degree (step 1), three cognitive ability tests (step 2), and two personality scales (step 3). Expectancies are computed for a 25% selection rate, with high performance defined as the top 25% of training grades.



Note: AR = Arithmetic Reasoning, RC = Reading Comprehension, BC = Block Counting, SDis = Self-Discipline, SUP = Stress Under Pressure (reverse coded).

Figure 7. Expectancy of high performance for alternate predictor combinations

1.7 Definitions of Technical Terms and Concepts

Bias. Systematic measurement error that differentially affects the scores of different groups of individuals.

Composite score. A total score that combines scores from several component tests according to a specified formula.

Confidence interval. A measure of uncertainty in an estimated value. Specifically, a range of scores within which the population value is expected to fall, with a specified level of certainty (e.g., 95%).

Concurrent validation design. A validity study where predictor and criterion scores are both obtained from incumbents at approximately the same time.

Contamination. System variance in scores that is irrelevant to the intended meaning of a measure.

Construct. A theoretical characteristics of an individual that is inferred from observed behavior or test scores.

Construct validity evidence. Evidence that test scores measure the intended theoretical characteristic.

Content validity evidence. Evidence based on expert judgement that the content of a test is representative of important work activities or work-related personal characteristics.

Convergent evidence. Evidence supporting the meaning of test scores based on the correlation with other measures of the same characteristic.

Correlation. A statistic reflecting the strength of linear relationship between two variables.

Criterion. An outcome variable valued by the organization that an assessment is attempting to predict, such as work performance, productivity, accident rate, or training performance.

Criterion-related validity evidence. Empirical relationship between scores on a predictor and scores on a criterion measure.

Cross-validation. When a scoring system or predictor weights are empirically derived in one sample, application of the system/weights to a different sample from the same population to investigate the stability of prediction.

Cutoff score. A score above which applicants are selected for further consideration in the selection process.

Deficiency. Failure of a measure to fully represent the intended theoretical domain.

Discriminant evidence. Evidence indicating whether two tests intended to measure different constructs are sufficiently uncorrelated to be considered two distinct constructs. Used together the convergent evidence to support construct validity.

Effect size. A statistical index of the strength of a relationships or group difference.

Imputation. A process for inferring plausible values for missing data.

Meta-analysis (a.k.a., validity generalization) A statistical procedure where results from several independent studies combined to estimate the relationship between variables.

Moderator variable. A variable that affects the strength, form, or direction of a predictor–criterion relationship.

Outlier. A value of a variable that is substantially different from the overall distribution of scores. Extreme outliers can have undue influence on statistical results, and should be carefully scrutinized.

Power. The probability that a statistical test will yield a significant result, if an effect of the specified magnitude indeed exists in the population.

Predictive validation design. A validity study where predictors scores are obtained from applicants and their criterion scores are obtained at a later point it time.

Reliability. The degree to which scores on a measure are consistent across potential sources of measurement error (e.g. time, raters, items). The reliability coefficient is a value between 0 and 1 indicating the degree to which scores are free from random measurement errors.

Restriction of range. Reduction in the variance of a sample relative to the full range of scores in the population of interest, resulting from incomplete sampling of participants for inclusion in the study. This is common in validation research because low-scoring individuals are not hired and therefore cannot be included in a validation study.

Shrinkage-adjusted R. An adjustment to the multiple correlation coefficient, accounting for the tendency of a regression model to fit a new sample less well than in the original sample on which the model was estimated.

Standard error (SE). A measure of uncertainty in an estimated value. Specifically, the magnitude of estimation errors to be expected due to sampling error.

Statistical significance. A result is inconsistent with the null hypothesis at a specified probability level, justifying rejection of the null hypothesis and conclusion that a relationship exists in the population.

Validity. The degree to which the accumulated evidence supports specific interpretations of scores and the proposed uses of a selection procedure.

2.0 REFERENCES

Aguinis, H., Werner, S., Lanza Abbott, J., Angert, C., Park, J. H., & Kohlhausen, D. (2010). Customer-centric science: Reporting significant research results with rigor, relevance, and practical impact in mind. *Organizational Research Methods*, *13*, 515-539.

Allred, L. J. (1991). Alternatives to the validity coefficient for reporting the test-criterion relationship. In National Research Council, *Performance Assessment for the Workplace, Volume II: Technical Issues* (pp. 158-206). Washington, DC: National Academies Press.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018).Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*, 3-25.

Arthur, W., Doverspike, D., Barrett, G. V., & Miguel, R. (2013). Chasing the Title VII holy grail: The pitfalls of guaranteeing adverse impact elimination. *Journal of Business and Psychology*, *28*(4), 473-485.

Beatty, A. S., Barratt, C. L., Berry, C. M., & Sackett, P. R. (2014). Testing the generalizability of indirect range restriction corrections. *Journal of Applied Psychology*, *99*, 587-598.

Bobko, P., Roth, P. L., & Bobko, C. (2001). Correcting the effect size of d for range restriction and unreliability. *Organizational Research Methods*, *4*(1), 46-61.

Bobko, P., Roth, P. L., & Buster, M. A. (2007). The usefulness of unit weights in creating composite scores: A literature review, application to content validity, and meta-analysis. *Organizational Research Methods, 10*(4), 689-709.

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology*, *100*(2), 431.

Cohen, D., Fortney, D., & Tison, E. (2017). Structuring a Traditional EEO Adverse Impact Analysis: The 2×2 Table. In S. B. Morris & E. M. Dunleavy (Eds.), *Adverse Impact Analysis: Understanding Data, Statistics and Risk* (pp. 49-70). NY: Routledge.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahway, NJ: Lawrence Erlbaum Associates.

Cucina, J. M., Berger, J. L., & Busciglio, H. H. (2017). Communicating criterion-related validity using expectancy charts: a new approach. *Personnel Assessment and Decisions*, *3*(1), 1-14.

Hattrup, K. (2012). Using composite predictors in personnel selection. In N. Schmitt (Ed.), *The Oxford handbook of personnel assessment and selection* (pp. 297-319). NY: Oxford University Press.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed). Hoboken, NJ: Wiley.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Sage.

Jeanneret, P. R., & Zedeck, S. (2010). Professional guidelines/standards. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (p. 593–625). Routledge/Taylor & Francis Group.

Jones, J. A., & Waller, N. G. (2013). Computing confidence intervals for standardized regression coefficients. *Psychological methods*, *18*(4), 435-453.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *17*, 137-152.

Kuncel, N. R., & Rigdon, J. (2013). Communicating research findings. In N. W. Schmitt, S. Highhouse, & I. B. Weiner (Eds.), *Handbook of psychology: Industrial and organizational psychology* (p. 43–58). Hoboken, NJ: John Wiley & Sons Inc.

LeBreton, J. M., Scherer, K. T., & James, L. R. (2014). Corrections for criterion reliability in validity generalization: A false prophet in a land of suspended judgment. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *7*(4), 478-500.

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational psychologist*, *38*(1), 43-52.

McDaniel, M. A., Kepes, S., & Banks, G. C. (2011). The Uniform Guidelines are a detriment to the field of personnel selection. *Industrial and Organizational Psychology*, *4*(4), 494-514.

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79*(4), 599.

Morris, S. B. (2017). Statistical Significance Testing in Adverse Impact Analysis. In S. B. Morris & E. M. Dunleavy (Eds.), *Adverse Impact Analysis: Understanding Data, Statistics and Risk* (pp. 71-91). NY: Routledge.

Morris, S. B., & Lobsenz, R. (2000). Significance tests and confidence intervals for the adverse impact ratio. *Personnel Psychology, 53*, 89-111.

Oswald, F. L., Dunleavy, E. M., & Shaw, A. (2017). Measuring Practical Significance in Adverse Impact Analysis. In S. B. Morris & E. M. Dunleavy (Eds.), *Adverse Impact Analysis: Understanding Data, Statistics and Risk* (pp. 92-112). NY: Routledge.

Pedhazur, E.J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach.* Mahwah, NJ: Erlbaum.

Ployhart, R. E. & Kubisiak, U. C. (Ed.). (2020). Briefing validation results. *Air Force Research Labs Best Practices Guides* (Institute Report #TBD). Wright-Patterson Air Force Base, OH. Airf Force Materiel Command, United States Air Force.

Ployhart, R. E., & Holtz, B. C. (2008). The diversity-validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology, 61*, 153-172.

Preacher, K. J., Curran, P. J., & Bauer, D. J. (2006). Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics*, *31*(4), 437-448.

Putka, D. J., Beatty, A. S., & Reeder, M. C. (2018). Modern prediction methods: New perspectives on a common problem. *Organizational Research Methods*, *21*(3), 689-732.

Raju, N. S., Bilgic, R., Edwards, J. E., & Fleer, P. F. (1997). Methodology review: Estimation of population validity and cross-validity, and the use of equal weights in prediction. *Applied Psychological Measurement*, *21*(4), 291-305.

Raju, N. S., Bilgic, R., Edwards, J. E., & Fleer, P. F. (1999). Accuracy of population validity and cross-validity estimation: An empirical comparison of formula-based, traditional empirical, and equal weights procedures. *Applied Psychological Measurement*, *23*(2), 99-115.

Roth, P. L., Bobko, P., & Switzer III, F. S. (2006). Modeling the behavior of the 4/5ths rule for determining adverse impact: Reasons for caution. *Journal of Applied Psychology*, *91*(3), 507.

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

Schmidt, F. (1996). Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for Training of Researchers. *Psychological Methods*, *1*, 115-129.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*, 262-274.

Schmitt, N. (2007). The value of personnel selection: Reflections on some remarkable claims. *Academy of Management Perspectives*, *21*(3), 19-23.

Schmitt, N., & Ployhart, R. E. (1999). Estimates of cross-validity for stepwise regression and with predictor selection. *Journal of Applied Psychology, 84*(1), 50.

Society for Industrial and Organizational Psychology (2018). *Principles for the Validation and Use of Personnel Selection Procedures* (5th ed.). *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.

Tonowski, R. (2017). Thoughts from an EEO Agency Perspective. In S. B. Morris & E. M. Dunleavy (Eds.), *Adverse Impact Analysis: Understanding Data, Statistics and Risk* (pp. 278-297). NY: Routledge.

Uniform guidelines on employee selection procedures (1978). 43 Fed. Reg. 38295 (August 25 1978); 29 CFR 1607.

U. S. Department of Labor, Employment and Training Administration (1999). *Testing and assessment: An employer's guide to good practices*. Retrieved from https://wdr.doleta.gov/opr/fulltext/document.cfm?docn=6032

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, *70*(2), 129-133.

Zhang, D. C. (2018). Utility of alternative effect size statistics and the development of a web-based calculator: Shiny-AESC. *Frontiers in Psychology*, *9*, 1221.

Zhang, D.C., Highhouse, S., Brooks, M.E. & Zhang, Y. (2018). Communicating the validity of structured job interviews with graphical visual aids. *International Journal of Selection and Assessment, 26*, 93-108.

LIST OF SYMBOLS, ABBREVIATIONS AND ACROYMS

AF-WIN	Air Force Work Interest Navigator
AFOCD	Air Force Officer Classification Directories
AFECD	Air Force Enlisted Classification Directories
AFOQT	Air Force Officer Qualifying Test
AFPC	Air Force Personnel Center
AFS	Air Force SpecialtyAFSC Air Force Specialty Code
AFHRL	Air Force Human Resources Laboratory
AIC	Akaike Information Criterion
AR	Arithmetic Reasoning
ASVAB	Armed Services Vocational Aptitude Battery
EDPT	Electronic Data Processing Test
EP	Expected number passing
LR	Likelihood ratio
MDPP	Multidimensional pairwise preference
MEPS	Military Entrance Processing Stations
METS	Military Entrance Test Sites
N	Number (sample size)
NP	Number passing
PCSM	Pilot Candidate Selection Method
RC	Reading Comprehension
SDI	Self-Description Inventory
SDis	Self-Discipline
SR	Selection ratio
SUP	Stress Under Pressure

r	Correlation
r^2	Coefficient of determination
R	Multiple Correlation
TAPAS	Tailored Adaptive Personality Assessment System
TBAS	Test of Basic Aviation Skills

Appendix A

The following table provides a summary of scientific and professional standards related to reporting findings from a criterion-related validation study. Standards related to other types of research contexts, other types validity evidence (content, construct), and other aspects of selection systems (e.g., test fairness) are not included. The wording is largely copied directly from the source material, with minor editing for readability. Where internal references to other parts of the standards were found, those are preserved, and the reader is directed to the original source for more information.

Summary of Professional Standards for Reporting Validation Research

Topic	Standards	SIOP Principles	Uniform Guidelines	Journal Article Reporting Standards
Intended Uses of Selection System	Standard 11.1. Prior to development and implementation of an employment or credentialing test, a clear statement of the intended interpretations of test scores for specified uses should be made. The subsequent validation effort should be designed to determine how well this has been achieved for all relevant subgroups.	In designing a validation effort, whether based on existing evidence, new evidence, or both, primary consideration should be given to the design features necessary to support the proposed uses. Examples of such features include the work to be targeted (e.g., one job title or job family), the relevant candidate pool (e.g., experienced or inexperienced candidates), the uniqueness of the operational setting (e.g., one homogeneous organization or many different organizations), and relevant criterion measures (e.g., performance or turnover). (p. 7)	(2) Problem and setting. An explicit definition of the purpose(s) of the study and the circumstances in which the study was conducted should be provided. A description of existing selection procedures and cutoff scores, if any, should be provided.	

Topic	Standards	SIOP Principles	Uniform Guidelines	Journal Article Reporting Standards
	Standard 11.10. If tests are to be used to make job classification decisions (e.g., if the pattern of predictor scores will be used to make differential job assignments), evidence that scores are linked to different levels or likelihoods of success among jobs, job groups, or job levels is needed.			
Variables - General		The estimates of predictor score reliability that are most appropriate in a given study will depend on the measurement design underlying one's predictor measures, the conditions of measurement one wishes to generalize scores across (e.g., raters, items, or occasions), and the ways in which the predictor measure will be used (e.g., for rank ordering applicants, or for making pass-fail or hire-no hire decisions; Haertel, 2006; Hunter & Schmidt, 1996; Putka & Sackett, 2010). When reporting estimates of predictor reliability, one should clearly describe the measurement design underlying underlying the collection of data on which indices of reliability are being estimated and clarify the sources of error	(7) Description of selection procedures. Any measure, combination of measures, or procedure studied should be completely and explicitly described or attached (essential). If commercially available selection procedures are studied, they should be described by title, form, and publisher (essential). Reports of reliability estimates and how they were established are desirable.	Define all primary and secondary measures and covariates, including measures collected but not included in the report.

Topic	Standards	SIOP Principles	Uniform Guidelines	Journal Article Reporting Standards
		that are reflected in the reported indices of reliability. (p. 13).		
			(8) Where revisions have been made in a selection procedure to assure compatibility between successful job performance and the probability of being selected, the studies underlying such revisions should be included (essential).	Describe methods used to enhance the quality of measurements, including training and reliability of data collectors and use of multiple observations
				Estimate and report values of reliability coefficients for the scores analyzed (i.e., the researcher's sample), if possible. Provide estimates of convergent and discriminant validity where relevant.

Topic	Standards	SIOP Principles	Uniform Guidelines	Journal Article Reporting Standards
				Report estimates related to the reliability of measures, including: interrater reliability for subjectively scored measures and ratings, test–retest coefficients in longitudinal studies in which the retest interval corresponds to the measurement schedule used in the study, internal consistency coefficients for composite scales in which these indices are appropriate for understanding the nature of the instruments being used in the study

Topic	Standards	SIOP Principles	Uniform Guidelines	Journal Article Reporting Standards
Variables - Criteria	Standard 1.17. When validation relies on evidence that test scores are related to one or more criterion variables, information about the suitability and technical quality of the criteria should be reported.	Criterion validation studies, when conducted, should report the following in detail: a description of the criterion measures; the rationale for their use; the data collection procedures; and a discussion of the measures' relevance, reliability, possible deficiencies, possible sources of contamination, and freedom from or control of biasing sources of variance. If the testing professional developed the criterion measure, then the report should include the rationale and steps taken to develop it, so it can be well understood and, if needed, replicated in future validation studies. (p. 34)	(5) Criterion measures. The bases for the selection of the criterion measures should be provided, together with references to the evidence considered in making the selection of criterion measures (essential). A full description of all criteria on which data were collected and means by which they were observed, recorded, evaluated, and quantified, should be provided (essential). If rating techniques are used as criterion measures, the appraisal form(s) and instructions to the rater(s) should be included as part of the validation evidence, or should be explicitly described and available (essential). All steps taken to insure that criterion measures are free from factors which would unfairly alter the scores of members of any group should be described (essential).	(see Variables - General)

Topic	Standards	SIOP Principles	Uniform Guidelines	Journal Article Reporting Standards
	Standard 11.7. When empirical evidence of predictor-criterion relationships in part of the pattern of evidence used to support test use, the criterion measure(s) used should reflect the criterion construct domain of interest to the organization. All criterion used should represent important work behaviors or work outputs, either on the job in the job-relevant training, as indicated by an appropriate review of information about the job.	The most appropriate estimate(s) of criterion reliability in a given study will depend on the measurement design underlying one's criterion measures, the conditions of measurement one wishes to generalize scores across, and the way in which the criterion measure will be used (Hunter & Schmidt, 1996; Putka & Hoffman, 2014; Putka & Sackett, 2010). When reporting estimates of criterion reliability, one should clearly describe the measurement design used and clarify what sources of error are reflected in the reported indices of reliability (e.g., rater-specific, item-specific, or occasion-specific errors). (p. 12)	(3) Job analysis or review of job information. A description of the procedure used to analyze the job or group of jobs, or to review the job information should be provided (Essential). Where a review of job information results in criteria which may be used without a full job analysis (see section 14B(3)), the basis for the selection of these criteria should be reported (Essential). Where a job analysis is required a complete description of the work behavior(s) or work outcome(s), and measures of their criticality or importance should be provided (Essential). The report should describe the basis on which the behavior(s) or outcome(s) were determined to be critical or important, such as the proportion of time spent on the respective behaviors, their level of difficulty, their frequency of performance, the consequences of error, or other appropriate factors (Essential).	

Topic	Standards	SIOP Principles	Uniform Guidelines	Journal Article Reporting Standards
	Standard 1.18. When it is asserted that a certain level of test performance predicts adequate or inadequate criterion performance, information about the levels of criterion performance associated with given levels of test scores should be provided.			
Predictor Scoring and Combination	Standard 1.19. If test scores are used in conjunction with other variables to predict some outcome or criterion, analyses based on statistical models of the prediction-criterion relationship should include those additional relevant variables along with the test scores.	Methods and algorithms used to score content should be fully described. For example, when weighted scores, derived scales, or composite or categorical scores are used, rationale should be provided in detail. When performance tasks, work samples, or other methods requiring some element of judgment are used, a description of the type of rater training conducted and scoring criteria should be provided. (p. 34)	(10) Uses and applications. The methods considered for use of the selection procedure (e.g., as a screening device with a cutoff score, for grouping or ranking, or combined with other procedures in a battery) and available evidence of their impact should be described (essential). This description should include the rationale for choosing the method for operational use, and the evidence of the validity and utility of the procedure as it is to be used (essential). The purpose for which the procedure is to be used (e.g., hiring, transfer, promotion) should be described (essential).	

Topic	Standards	SIOP Principles	Uniform Guidelines	Journal Article Reporting Standards
		The recommendations for implementation of selection procedures and the rationale supporting the recommendations (e.g., the use of rank ordering, score bands, or cutoff scores, and the means of combining information in making personnel decisions) should be provided. (p. 35)	(10) If weights are assigned to different parts of the selection procedure, these weights and the validity of the weighted composite should be reported (essential).	
			(10) If the selection procedure is used with a cutoff score, the user should describe the way in which normal expectations of proficiency within the work force were determined and the way in which the cutoff score was determined (essential).	

Topic	Standards	SIOP Principles	Uniform Guidelines	Journal Article Reporting Standards
Description of sample	Standard 1.8. The composition of any sample of test takers from which validity evidence is obtained should be described in as much detail as is practical and permissible, including major relevant socio-demographic and developmental characteristics.	The sampling procedure and the characteristics of the research sample relative to the appropriate interpretation of the results should be described. The description should include a definition of the population that the sample is designed to represent, sampling biases that may detract from the representativeness of the sample, the significance of any deviations from representativeness for the interpretation of the results, and any statistical power analysis results. Data informing the potential restriction in the range of scores on predictors or criterion measures are especially important. (p. 34)	(6) Sample description. A description of how the research sample was identified and selected should be included (essential). The race, sex, and ethnic composition of the sample, including those groups set forth in section 4A above, should be described (essential). This description should include the size of each subgroup (essential). A description of how the research sample compares with the relevant labor market or work force, the method by which the relevant labor market or work force was defined, and a discussion of the likely effects on validity of differences between the sample and the relevant labor market or work force, are also desirable. Descriptions of educational levels, length of service, and age are also desirable.	Report major demographic characteristics (e.g., age, sex, ethnicity, socioeconomic status) and important topic-specific characteristics (e.g., achievement level in studies of educational interventions).
		Test developers should make clear whether psychometrics in the technical report refer to candidates or incumbents, and results for concurrent validation studies should not be represented as the results for predictive validation studies. (p. 34)	(3) Where two or more jobs are grouped for a validity study, job analysis information should be provided for each of the jobs, and the justification for the grouping (see section 14B(1)) should be provided (Essential).	Report inclusion and exclusion criteria, including any restrictions based on demographic characteristics.

Topic	Standards	SIOP Principles	Uniform Guidelines	Journal Article Reporting Standards
				Describe procedures for selecting participants, including sampling method if a systematic sampling plan was implemented, and percentage of sample approached that actually participated. Describe settings and locations where data were collected as well as dates of data collection. Describe agreements and payments made to participants. Describe institutional review board agreements, ethical standards met, and safety monitoring.
				Describe the sample size, power, and precision, including: intended sample size, achieved sample size, if different from the intended sample size, and determination of sample size (e.g., power analysis)
				Report the flow of participants, including total number of participants at each stage of the study

Topic	Standards	SIOP Principles	Uniform Guidelines	Journal Article Reporting Standards
Study Design	Standard 1.10. When validity evidence includes statistical analyses of test results, either along or together with data on other variables, the conditions under which the data were collected should be described in enough detail that users can judge the relevance of the statistical findings to local conditions. Attention should be drawn to any features of a validation data collection that are likely to differ from typical operational testing conditions and that could plausibly influence test performance	Reports of validation efforts should include enough detail to enable a testing professional competent in personnel selection to know what was done, draw independent conclusions in evaluating the research, replicate the study, and make recommendations regarding the use of the selection procedure. (p. 33)	(1) User(s), location(s), and date(s) of study. Dates and location(s) of the job analysis or review of job information, the date(s) and location(s) of the administration of the selection procedures and collection of criterion data, and the time between collection of data on selection procedures and criterion measures should be provided (Essential). If the study was conducted at several locations, the address of each location, including city and State, should be shown.	Describe methods used to collect data.
	Standard 11.8. Individuals conducting and interpreting empirical studies of predictor-criterion relationships should identify artifacts that may have influenced study findings, such as errors of measurement, range restriction, criterion deficiency, criterion contamination, and missing data. Evidence of the presence or absence of such features, and of actions taken to remove or control their influence, should be documented and made available as needed.			

Topic	Standards	SIOP Principles	Uniform Guidelines	Journal Article Reporting Standards
Data Cleaning		Testing professionals should also check their data for both univariate and multivariate outliers (Aguinis, Gottfredson, & Joo, 2013). Documentation should include how outliers were defined and identified. If clear outliers are found, sensitivity analyses should be performed to evaluate the effects of including and excluding outliers on the validation study results, or robust estimation/analysis techniques should be used that account for the presence of outliers. (p. 31)		Describe planned data diagnostics, including: criteria for post-data-collection exclusion of participants, if any, criteria for deciding when to infer missing data and methods used for imputation of missing data, definition and processing of statistical outliers, analyses of data distributions, data transformations to be used, if any
		Orr, Sackett, and DuBois (1991) report that most testing professionals oppose dropping outliers unless there is evidence that the data point is erroneous. Dropping outliers to obtain more favorable results is not appropriate. (p. 31)		

Topic	Standards	SIOP Principles	Uniform Guidelines	Journal Article Reporting Standards
Descriptive Statistics		Most data analyses will begin with descriptive statistics for predictor and criterion variables that present analyses of frequencies, central tendencies, and variances. Such descriptions should be provided for the total group and for relevant subgroups if they are large enough to yield reasonably reliable estimates. (p. 31)	(8) Measures of central tendency (e.g., means) and measures of dispersion (e.g., standard deviations and ranges) for all selection procedures and all criteria should be reported for each race, sex, and ethnic group which constitutes a significant factor in the relevant labor market (essential).	
Data Analysis - General			(8) Methods used in analyzing data should be described (essential).	Provide information detailing the statistical and data-analytic methods used.
				Report other data analyses performed, including adjusted analyses, if performed, indicating those that were planned and those that were not planned (though not necessarily in the level of detail of primary analyses).

Topic	Standards	SIOP Principles	Uniform Guidelines	Journal Article Reporting Standards
Data Analysis - Missing Data		When there are missing data, the testing professional should provide (a) a summary of missing data patterns and the nature of the missingness (e.g., missing at random, missing completely at random, missing not at random) and (b) justification for the missing data technique adopted for analyses. (p. 31)		Provide information on missing data, including the frequency or percentages of missing data, empirical evidence and/or theoretical arguments for the causes of data that are missing (for example, missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR)), and methods actually used for addressing missing data, if any.
		Most data analyses will begin with descriptive statistics for predictor and criterion variables that present analyses of frequencies, central tendencies, and variances. Such descriptions should be provided for the total group and for relevant subgroups if they are large enough to yield reasonably reliable estimates. (p. 31)		Provide descriptions of each primary and secondary outcome, including the total sample and each subgroup that includes the number of cases, cell means, standard deviations, and other measures that characterize the data used.

Topic	Standards	SIOP Principles	Uniform Guidelines	Journal Article Reporting Standards
Data Analysis - Statistical Significance, Effect Size and Confidence Intervals	Standard 1.20. When effect size measures (e.g., correlation between test scores and criterion measures, standardized mean test score differences between subgroups) are used to draw inference that go beyond describing the sample or samples on which data have been collected, indices of the degree of uncertainty associate with these measures (e.g., standard errors, confidence intervals, or significance tests) should be reported.		(8) Statements regarding the statistical significance of results should be made (essential).	Report results of all inferential tests conducted, including exact p values if null hypothesis significance testing (NHST) methods were used, and reporting the minimally sufficient set of statistics (e.g., dfs, mean square [MS] effect, MS error) needed to construct the tests
				Report effect-size estimates and confidence intervals on estimates that correspond to each inferential test conducted, when possible

Topic	Standards	SIOP Principles	Uniform Guidelines	Journal Article Reporting Standards
Data Analysis – Statistical Methods & Assumptions				For complex data analyses (for example, structural equation modeling analyses, hierarchical linear models, factor analysis, multivariate analyses, and so forth), provide the details of the models estimated, associated variance–covariance (or correlation) matrix or matrices, and identification of the statistical software used to run the analyses (e.g., SAS PROC GLM or the particular R package)
				Report estimation problems (e.g., failure to converge, bad solution spaces), regression diagnostics, or analytic anomalies that were detected and solutions to those problems.
				Report any problems with statistical assumptions and/or data distributions that could affect the validity of findings.

Topic	Standards	SIOP Principles	Uniform Guidelines	Journal Article Reporting Standards
Presenting Results - General		The reports must accurately portray the findings, as well as the interpretations of and decisions based on the results. Research findings that qualify the conclusions or support the generalizability of results should be reported. (P. 33)		Provide a clear differentiation between primary hypotheses and their tests–estimates, secondary hypotheses and their tests–estimates, and exploratory hypotheses and their test–estimates
		Research reports or administration manuals should help readers make appropriate interpretations of data and should warn them against common misuses of information. (p. 35)		
Presenting Results - Comprehensiveness		All summary statistics that relate to the conclusions drawn by the testing professional and the recommendations for use should be included. Complete statistical results related to the development and validation, not just statistically significant or supportive results, should be presented and clearly labeled. (p. 34)	(8) The magnitude and direction of all relationships between selection procedures and criterion measures investigated should be reported for each relevant race, sex, and ethnic group and for the total group (essential). Where groups are too small to obtain reliable evidence of the magnitude of the relationship, need not be reported separately.	

Topic	Standards	SIOP Principles	Uniform Guidelines	Journal Article Reporting Standards
Presenting Results - Statistical Adjustments	Standard 1.21. When statistical adjustments, such as those for restriction of range or attenuation, are made, both adjusted and unadjusted coefficients, as well as the specific procedure used, and all statistics used in the adjustment, should be reported. Estimates of the construct-criterion relationship that remove the effects of measurement error on the test should be clearly reported as adjusted estimates.	Both uncorrected and corrected values should be presented when corrections are made for statistical artifacts such as restriction of range or unreliability of the criterion. (p. 34)	(8) Any statistical adjustments, such as for less than perfect reliability or for restriction of score range in the selection procedure or criterion should be described and explained; and uncorrected correlation coefficients should also be shown (essential).	
			(8) Where the statistical technique categorizes continuous data, such as biserial correlation and the phi coefficient, the categories and the bases on which they were determined should be described and explained (essential).	

Appendix B

R code for examples presented in Best Practices for Briefing Validation Results

All code assumes the data are located in a dataframe named *dat*. The examples are based on a hypothetical validation study with 200 cases the following variables:

Degree	Prior degree in related field
AR	Arithmetic Reasoning Test
RC	Reading Comprehension Test
BC	Block Counting Test
SDis	Self-Discipline
SUP	Stress Under Pressure (Reverse Scored)
Performance	Training course grade (% of total)
Pass	Passing grade in training course (0=fail, 1=pass)
Minority	Examinee minority status (0=non-minority, 1=minority)

Click the paper clip on the panel to the left to open the simulated dataset “sample data.csv.” To load the data, copy the csv file to the working directory, then run the following commands

```
dat <- read.csv("sample data.csv")
N <- nrow(dat)
View(dat)
```

```
# Example 1 -------------------------------------------------------------

# Table of Validities
predictorList <- list("AR","RC","BC","SDis","SUP")
corStats <- lapply(predictorList,function(x) cor.test(dat[,x],dat[,"Performance"]))
validity <- sapply(1:length(predictorList),function(x) corStats[[x]]$estimate)
pvalue <- sapply(1:length(predictorList),function(x)corStats[[x]]$p.value)
validityTable <- data.frame(Predictor=unlist(predictorList),validity,pvalue)

# create a string variable with correlation rounded to 2 digits with significance stars
mystars <- ifelse(validityTable$pvalue < .001, "***",ifelse(validityTable$pvalue < .01, "**",
ifelse(validityTable$pvalue < .05, "* ", "  ")))
validityTable$r <- sprintf("%.2f%s",validityTable$validity,mystars)
validityTable$Predictor <- c("Arithmetic Reasoning", "Reading Comprehension","Block
Counting",
"Self-Discipline","Stress Under Pressure")

# Create a table with confidence intervals
validityCI <- sapply(1:length(predictorList),function(x) round(corStats[[x]]$conf.in,2))
validityCI <- sapply(1:length(predictorList), function(x) paste0("[",validityCI[1,x],",
",validityCI[2,x],"]"))
validityTableCI <-
data.frame(Predictor=validityTable$Predictor,Validity=validityTable$r,CI=validityCI)
```

```
write.table(validityTableCI, "clipboard", sep="\t", row.names=TRUE,col.names = NA)
# after writing to clipboard, paste into excel

# Create a table with standard errors
# Note: the cor.table() procedure does not output the SE, but it can be derived from the t-test.
# Since t = r/SE, SE = r/t
validitySE <- sapply(1:length(predictorList),function(x)
corStats[[x]]$estimate/corStats[[x]]$statistic)
validitySE <- paste0("(",round(validitySE,2),")")
validityTableSE <-
data.frame(Predictor=validityTable$Predictor,Validity=validityTable$r,SE=validitySE)
write.table(validityTableSE, "clipboard", sep="\t", row.names=TRUE,col.names = NA)
# after writing to clipboard, paste into excel

# Example 2 -------------------------------------------------------------

# Formatted correlation matrix
# ----------------------------------------------------
# Function to produce correlation table with significance stars
# Adapted from http://myowelt.blogspot.com/2008/04/beautiful-correlation-tables-in-r.html
corstars <- function(x){
  require(Hmisc)
  x <- as.matrix(x)
  R <- rcorr(x)$r
  p <- rcorr(x)$P

  # define significance flags
  mystars <- ifelse(p < .001, "***",ifelse(p < .01, "**", ifelse(p < .05, "* ", "  ")))

  # Round and truncate correlations to 2 decimal places
  R <- format(round(cbind(rep(-1.11, ncol(x)), R), 2))[,-1]

  # Combined correlations with stars
  Rnew <- matrix(paste(R, mystars, sep=""), ncol=ncol(x))
  diag(Rnew) <- paste(diag(R), "  ", sep="")
  rownames(Rnew) <- colnames(x)
  colnames(Rnew) <- paste(colnames(x), "", sep="")

  # Remove upper triangle
  Rnew[upper.tri(Rnew)] <- ""

  # Output formatted correlation matrix
  Rnew <- as.data.frame(Rnew)
  return(Rnew)
  }

# Create correlation matrix for the current data
```

```
dataMain <- subset(dat, select = c(AR,RC,BC,SDis,SUP,Performance))
M <- round(sapply(dataMain, mean),2)
SD <- round(sapply(dataMain, sd),2)
rMatrixTable <- corstars(as.matrix(dataMain))
rMatrixTable <- data.frame(M,SD,rMatrixTable)

# Copy table to clipboard, then paste into Excel
write.table(rMatrixTable, "clipboard", sep="\t", row.names=TRUE,col.names = NA)

# Example 3 ------------------------------------------------------------

# Comprehensive regression table
library(psychometric) # provides function to compute confidence interval on R2
library(QuantPsyc) # provides lm.beta function to compute standardized coefficients
library(data.table) # provides rbindlist function

regF <- lm(data = dat, Performance ~ AR + RC + BC + SDis + SUP)
regF.out <- summary(regF)  # standard output from lm()
regF.b <- regF.out$coefficients  # regression coefficients and t-test
regF.b[,1:3] <- round(regF.b[,1:3],2)
regf.Beta <- round(lm.beta(regF),2)  # standardized coefficients
regF.CI <- round(confint(regF),2)    # confidence interval for raw coefficients
regF.Rsq <- CI.Rsqlm(regF)[1:2]  # Rsq and its confidence interval
regF.R <- (sqrt(regF.Rsq[1])) # Multiple R
regF.sigma <- round(regF.out$sigma,2)
regF.F <- summary(regF)$fstatistic
regF.Fdf <- paste(regF.F[2],regF.F[3], sep = ", ")
regF.F <- round(regF.F[1],2)
regF.pF <- anova(lm(data = dat, Performance ~ 1),regF)[6][2,1]

# Round significance level and truncate below .001 for readability
pOut <- function(x) ifelse(x < .001, "< .001",as.character(round(x,3)))
regF.p <- pOut(regF.b[,4])
regF.pF <- pOut(regF.pF)

# Estimated cross validity (adjusted R)
# See Raju, N.S., Bilgic, R., Edwards, J.E., & Fleer, P.F. (1997). Methodology review:
Estimation of
# population validity and cross-validity, and the use of equal weights in prediction. Applied
Psychological
# Measurement, 21(4), 291-305.
# Because the Browne formula estimates Rsq, and we take the square root to obtain the Multiple
R.
# Functions to apply Burket and Browne formulas.
BurketR <- function (n, k, R) (n * R^2 - k)/(R*(n-k))
BrowneR <- function (n, k, R) sqrt(((n - k - 3)*R^4 + R^2)/((n - 2*k - 2)*R^2 + k))
```

```
# Calculate adjusted R for current data
nPredictors <- 40  # specified total number of predictors considered
adjusted.R.Burket <- unname(round(BurketR(N, nPredictors, regF.R[1]),2))
adjusted.R.Browne <- unname(round(BrowneR(N, nPredictors, regF.R[1]),2))

# Round Rsq stats for output
regF.Rsq <- round(regF.Rsq, 2)
regF.R <- round(regF.R, 2)

# Combine results into single table
colnames(regF.Rsq) <- c("Coeff","SE")
colnames(regF.R) <- colnames(adjusted.R.Burket) <- "Coeff"
#colnames(regF.CI) <- c("LCL","UCL")

regTable <- data.frame(Coeff = regF.b[,1],SE=regF.b[,2], Beta = c("",regf.Beta), t = regF.b[,3],
p = regF.p)
regTable <-
list(regTable,regF.Rsq,regF.R,adjusted.R.Burket,list(regF.sigma),list(regF.F),list(regF.Fdf),
        list(regF.pF))
regTable <- rbindlist(regTable, fill = TRUE)

predictorLabels <- variable.names(regF)
rownames(regTable) <- c(predictorLabels,"Rsq","Multiple R", "adj. R","Residual
SD","F","df","p")

write.table(regTable, "clipboard", sep = "\t", na = "",col.names = NA)
# after copying to clipboard, past into excel

# Example 4 ------------------------------------------------------------

# Concise regression table with standardized coefficients
library(psychometric) # provides function to compute confidence interval on R2
library(QuantPsyc) # provides lm.beta function to compute standardized coefficients
library(data.table) # provides rbindlist function

regF <- lm(data = dat, Performance ~ AR + RC + BC + SDis + SUP)
regF.coeff <- coef(regF)    # raw regression coefficients
regF.CI <- confint(regF)    # confidence interval for raw coefficients
regF.Beta <- lm.beta(regF)  # standardized coefficients
regF.p <- summary(regF)$coefficients[,4]
regF.pStars <-ifelse(regF.p < .01, "***",ifelse(regF.p < .01, "**", ifelse(regF.p < .05, "* ", "
")))
regF.Rsq <- CI.Rsqlm(regF)  # Rsq and its confidence interval
regF.R <- round(sqrt(regF.Rsq)[-2],2) # Multiple R
regF.Rsq <- round(regF.Rsq,2)
regF.F <- summary(regF)$fstatistic
regF.pF <- anova(lm(data = dat, Performance ~ 1),regF)[6][2,1]
```

```
# Compute confidence interval for standardized coefficients
# by standardizing endpoints of CI for raw coefficiencts
sx <- sapply(regF$model[-1],sd)  # sd for predictors
sy <- sd(regF$model[[1]])        # sd for criterion
regF.Beta.CI <- regF.CI[-1,]*sx/sy
regF.Beta <- round(data.frame(regF.Beta,regF.Beta.CI),2)
names(regF.Beta) <- c("Coeff","CI.lower","CI.upper")

# Estimated cross validity (adjusted R)
# See Raju, N.S., Bilgic, R., Edwards, J.E., & Fleer, P.F. (1997). Methodology review:
Estimation of
# population validity and cross-validity, and the use of equal weights in prediction. Applied
Psychological
# Measurement, 21(4), 291-305.
# Note that Browne formula estimates Rsq; we take the square root to obtain the Multiple R.
# Functions to apply Burket and Browne formulas.
BurketR <- function (n, k, R) (n * R^2 - k)/(R*(n-k))
BrowneR <- function (n, k, R) sqrt(((n - k - 3)*R^4 + R^2)/((n - 2*k - 2)*R^2 + k))

# Calculate adjusted R for current data
nPredictors <- 40
adjusted.R.Burket <- round(BurketR(N, nPredictors, regF.R[1]),2)
adjusted.R.Browne <- round(BrowneR(N, nPredictors, regF.R[1]),2)

# Table regression results with standardized coefficients
predictorLabels <- variable.names(regF)
CIout <- sprintf("[%.2f, %.2f]", regF.Beta$CI.lower, regF.Beta$CI.upper)
BetaOut <- data.frame(Coeff=paste0(regF.Beta[,1],regF.pStars[-1]),CI=CIout)
Rout <- data.frame(Coeff = unname(regF.R[1]), CI = sprintf("[%.2f,
%.2f]",regF.R[2],regF.R[3]))
Rout <- rbind(Rout,data.frame(Coeff=unname(adjusted.R.Burket),CI=""))
regTableBeta <- rbindlist(list(BetaOut,Rout))
rownames(regTableBeta) <- c(predictorLabels[-1],"Multiple R","Adjusted R (Burket)")
#regTableBeta

# Copy table to clipboard
write.table(regTableBeta,"clipboard", sep = "\t",col.names = NA)

# Example 5 ---------------------------------------------------------------

# Table for Hierarchical Regression
# This code assumes predictors are in the same order in all models and the liast model contains
all predictors
# Adapted from https://thomasleeper.com/Rcourse/Tutorials/wordoutput.html
# Conduct a sequence of hierarchically nested models, assigning each output to a different
object.
# Then combine into a list
regM1 <- lm(data = dat, Performance ~ Degree)
regM2 <- lm(data = dat, Performance ~ Degree + AR + RC + BC )
regM3 <- lm(data = dat, Performance ~ Degree + AR + RC + BC + SDis + SUP)
regOut <- list(regM1,regM2,regM3) # create list with all results
nModels <- length(regOut)

# regression coefficients
predictorLabels <- variable.names(regOut[[length(regOut)]])
s <- lapply(regOut,summary)
b <- lapply(s,function(x) round(coef(x),2))
nCoeff <- lapply(b,nrow)
regF.p <- lapply(s, function(x) x$coefficients[,4])
regF.pStars <- lapply(regF.p, function(x) ifelse(x < .01, "***",ifelse(x < .01, "**", ifelse(x <
.05, "* ", "  ")))) # significance flags

# formatted coefficients
bOut <- sapply(b, function(x) sprintf("%.2f (%.2f)",x[,1],x[,2]))
bOut <- mapply(function(x,y) paste0(x,y),bOut,regF.pStars)

# model summary statistics
sigma <- sapply(s, function(x) round(x$sigma, 2))
Rsq <- sapply(s, function(x) round(c( x$r.squared), 2))
R <- round(sqrt(Rsq),2)
Ftest <- sapply(s, function(x) round(x$fstatistic,2))
Fsig <- pf(Ftest[1,],Ftest[2,],Ftest[3,],lower.tail = FALSE)
pOut <- function(x) ifelse(x < .001, "< .001",as.character(round(x,3)))
Fsig <- pOut(Fsig)
Fdf <- paste(Ftest[2,],Ftest[3,], sep=", ")
Ftest <- round(Ftest[1,],2)

# model comparison stats
deltaRsq <- deltaF <- array(NA,dim = nModels)
for(i in 2:nModels) deltaRsq[i] <- Rsq[i]-Rsq[(i-1)]
anovaOut <- do.call(anova,regOut)
deltaF <- round(anovaOut$F,2)
deltaFp <- anovaOut$"Pr(>F)"
deltaFp <- pOut(deltaFp)

# combine all result in table
maxCoeff <- max(unlist(nCoeff))
nBlanks <- sapply(nCoeff, function(x) maxCoeff - x)
bOut <- mapply(function(x,y) c(x,rep("",y)),bOut,nBlanks)
HRegTable <- rbind(bOut, sigma, Rsq, R, Ftest, Fdf, Fsig, deltaRsq, deltaF, deltaFp)
```

```
colnames(HRegTable) <- sapply(seq(1,nModels), function(x) paste("Model",x)) # create column label
rownames(HRegTable) <- c(predictorLabels, "Residual SD", "R-Squared", "Multiple R", "F", "df","p","Delta Rsq", "Delta F", "p")
HRegTable

write.table(HRegTable, "clipboard", sep = '\t',col.names = NA, na = "")  # copy to clipboard
# After writing to clipboard, paste into Excel

# Example 6 ---------------------------------------------------------------

# Logistic Regression
library(rsq)
library(data.table)

lrMod <- glm(Pass ~ AR + RC + BC + SDis + SUP, data = dat, family = "binomial")
lrOut <- summary(lrMod)
lr.coeff <- lrOut$coefficients
lr.AIC <- round(lrOut$aic,2)
lr.CI <- confint(lrMod)
lr.b <- cbind(Coeff = coef(lrMod),confint(lrMod))
lr.OR <- exp(lr.b)
lr.LRtest <- round(anova(glm(Pass ~ 1, data = dat, family = "binomial"),lrMod),2)
lr.LRdf <- lr.LRtest$Df[2]
lr.LRtest <- lr.LRtest$Deviance[2]
lr.LRp <-  pchisq(lr.LRtest, lr.LRdf, lower.tail = FALSE)
lr.Rsq <- round(rsq(lrMod, type = "n"),2) # Negelkerke Rsq

# Summary table
pOut <- function(x) ifelse(x < .001, "< .001",as.character(round(x,3)))
lr.coeff[,1:3] <- round(lr.coeff[,1:3],2)
lr.coeff[,4] <- pOut(lr.coeff[,4])
lr.LRp <- pOut(lr.LRp)
stats <- list(rbind(lr.Rsq,lr.AIC,lr.LRtest,lr.LRdf,lr.LRp))
names(stats) <-  "Estimate"
lrTable <- rbindlist(list(as.data.frame(lr.coeff),stats), fill = TRUE)
rownames(lrTable) <- c(rownames(lr.coeff),"Negelkerke Rsq","AIC","LR Chi-sq","df","p")

write.table(lrTable, "clipboard", sep = '\t',col.names = NA, na = "")  # copy to clipboard
# after writing to clipboard, copy into Excel

# Predicted probability
predictorLabels <- rownames(lr.coeff)[-1]
focalPredictor <- "SUP"
predictorMeans <- sapply(dat[predictorLabels], mean)
focalRange <- sd(dat[,focalPredictor])*3
focalPredictorLevels <- seq(from=(predictorMeans[focalPredictor]-focalRange),
        to=predictorMeans[focalPredictor]+focalRange,length.out = 20)
otherPredictors <- predictorMeans[!names(predictorMeans) %in% focalPredictor]
predictorLevels <- data.frame(focalPredictorLevels,array(otherPredictors, dim =
c(1,length(otherPredictors))))
colnames(predictorLevels) <- c(focalPredictor,names(otherPredictors))

predProb <- predict(lrMod, newdata = predictorLevels, type = "response")
predictorLevels <- data.frame(predictorLevels,predProb)

# Plot results
library(ggplot2)
ggplot(data = predictorLevels, aes(x=predictorLevels[,focalPredictor], y=predProb)) +
        geom_line() +
        ylim(0,1) +
        xlab(focalPredictor) +
        ylab("Probability") +
        theme_bw()
```

# Example 7 --

```
# Table of validities with color coding
library(formattable)
predictorList <- list("AR","RC","BC","SDis","SUP")
corStats <- lapply(predictorList,function(x) cor.test(dat[,x],dat[,"Performance"]))
validity <- sapply(1:length(predictorList),function(x) corStats[[x]]$estimate)
pvalue <- sapply(1:length(predictorList),function(x)corStats[[x]]$p.value)
validityTable <- data.frame(Predictor=unlist(predictorList),validity,pvalue)

# create a string variable with correlation and significance stars
mystars <- ifelse(validityTable$pvalue < .001, "***",ifelse(validityTable$pvalue < .01, "**",
        ifelse(validityTable$pvalue < .05, "* ", "  ")))
validityTable$r <- sprintf("%.2f%s",validityTable$validity,mystars)  # round and add
significance stars
validityTable$Predictor <- c("Arithmetic Reasoning", "Reading Comprehension","Block
Counting",
        "Self-Discipline","Stress Under Pressure")

formattable(validityTable, align=c("l","c"), list(
  'validity' = FALSE, 'pvalue' = FALSE,
  'r' = formatter("span", style = x ~ ifelse(validity < .11, "background-color:NA",
      ifelse(validity<.21,"background-color:#FFFF99",
      ifelse(validity<.35,"background-color:#CCFFCC",
      "background-color:#80FF00"))))
))
# This code creates a graphic object. Save as image file and then insert as image into word or
powerpoint.
```

```
# Example 8 ------------------------------------------------------------

# Table of validity with bullet chart
# Bar indicates correlation coefficient
# Colored regions represent DOC categories:
# Red r<.11 "Unlikely to be useful", Yellow .11-.20 "Depends on Circumstances", Green .21 +
"Likely to be Useful" or "Very Beneficial"
library(sparkline)
library(formattable)

predictorList <- list("AR","RC","BC","SDis","SUP")
corStats <- lapply(predictorList,function(x) cor.test(dat[,x],dat[,"Performance"]))
validity <- sapply(1:length(predictorList),function(x) corStats[[x]]$estimate)
pvalue <- sapply(1:length(predictorList),function(x)corStats[[x]]$p.value)
validityTable <- data.frame(Predictor=unlist(predictorList),validity,pvalue)

# create a string variable with correlation and significance stars
mystars <- ifelse(validityTable$pvalue < .001, "***",ifelse(validityTable$pvalue < .01, "**",
ifelse(validityTable$pvalue < .05, "* ", "  ")))
validityTable$r <- sprintf("%.2f%s",validityTable$validity,mystars)  # round and add
significance stars
validityTable$Predictor <- c("Arithmetic Reasoning", "Reading Comprehension","Block
Counting","
          Self-Discipline","Stress Under Pressure")

customRed <- "#FF9999"
customYellow <- "#FFFF99"
customGreen1 <- "#CCFFCC"
customGreen2 <- "#80FF00"
validityTable$" " <- sapply(validityTable$validity, FUN=function(x)
as.character(htmltools::as.tags(
  sparkline(c(NA,as.numeric(x),max(x,.5),.21,.11), type = "bullet", performanceColor = "black",
          rangeColors = c(customGreen2,customYellow,customRed)))))
out <- as.htmlwidget(formattable(validityTable, align = c("l","c","c"), list("validity" = FALSE,
"pvalue" = FALSE)))
out$dependencies <- c(out$dependencies, htmlwidgets:::widget_dependencies("sparkline",
"sparkline"))
out
# This code creates a graphic object. Save as image file and then insert as image into word or
powerpoint.

# Example 9 ------------------------------------------------------------

# Compute expectancy chart for multiple predictors
# Note: The script for the Expectancyfunc() function must be located in the working directory.
library(Hmisc)  # provides cor.test function
library(ggplot2)  # graphics package
```

```
source("Cucina Expectancy Function.R") # R script for Expectancyfunc ()

# specify which predictors and criterion variable
predictorList <- list("AR","RC","BC","SDis","SUP")
criterionVariable <- "Performance"
corStats <- lapply(predictorList,function(x) cor.test(dat[,x],dat[,criterionVariable]))
validity <- sapply(1:5,function(x) corStats[[x]]$estimate)
names(validity) <- predictorList
validityCI <- t(sapply(1:5,function(x) corStats[[x]]$conf.int))

#Specify passing rate and success rate.
#These are convered into standardized cut scores.
PassRate <- .30
SuccessRate <- .60
PredictorCutoff <- qnorm(1-PassRate)
CriterionCutoff <- qnorm(1-SuccessRate)

# Expectancy for those who pass.
# Confidence interval computed using upper and lower bounds of confidence interval on valdity
expectancyPass <- sapply(validity, function (x)
Expectancyfunc(x,PredictorCutoff,Inf,CriterionCutoff,Inf)$expectancy*100)
expPassLCL <- sapply(validityCI[,1], function (x)
Expectancyfunc(x,PredictorCutoff,Inf,CriterionCutoff,Inf)$expectancy*100)
expPassUCL <- sapply(validityCI[,2], function (x)
Expectancyfunc(x,PredictorCutoff,Inf,CriterionCutoff,Inf)$expectancy*100)

#Expectancy for those who fail
expectancyFail <- sapply(validity, function (x) Expectancyfunc(x,-
Inf,PredictorCutoff,CriterionCutoff,Inf)$expectancy*100)
expFailLCL <- sapply(validityCI[,1], function (x) Expectancyfunc(x,-
Inf,PredictorCutoff,CriterionCutoff,Inf)$expectancy*100)
expFailUCL <- sapply(validityCI[,2], function (x) Expectancyfunc(x,-
Inf,PredictorCutoff,CriterionCutoff,Inf)$expectancy*100)

# Combine results into dataframe
out <-
data.frame(Test=unlist(predictorList),Range="Fail",Expectancy=expectancyFail,ExpLCL=expF
ailLCL,ExpUCL=expFailUCL)
out <-
rbind(out,data.frame(Test=unlist(predictorList),Range="Pass",Expectancy=expectancyPass,Exp
LCL=expPassLCL,ExpUCL=expPassUCL))
out$Test <- factor(out$Test, levels = c("AR","RC","BC","SDis","SUP"), labels = c("Arithmetic
\nReasoning","Reading Comp","Block Counting","Self Discipline","Stress Under \nPressure"))
View(out)

# Crate graph
```

```
ggplot(data = out, aes(x=Test, y=Expectancy, fill=Range)) +
  geom_bar(position = position_dodge(),stat = "identity")+
  geom_errorbar(aes(ymin=ExpLCL, ymax=ExpUCL), width=.3,position = position_dodge(.9))
+
  xlab(NULL) +
  ylab("Percent Successful") +
  labs(fill="Test Range") +
  coord_flip()
# Save plot as an image file or copy to clipboard, then insert into word or powerpoint.
```

# Example 10 ---

```
# Plot expectancy for multiple score ranges
# Produces both vertical and horizontal bar charts
library(Hmisc)  # provides cor.test function
library(ggplot2)  # graphics package
source("Cucina Expectancy Function.R") # R script for Expectancyfunc ()

# Compute validity and confidence interval for selected predictor
predictorVariable <- "AR"  # name of variable in dataframe
criterionVariable <- "Performance"
predictorLabel <- "Arithmetic Reasoning"  # Lable for plot title
rStats <- cor.test(dat[,predictorVariable],dat[,criterionVariable])
Validity <- rStats$estimate
ValidityCI <- rStats$conf.int

# Specify success rate, which is converted into a standardized cutoff
SuccessRate <- .25
CriterionCutoff <- qnorm(1-SuccessRate)

# Specify desired number of score ranges for predictor
nRanges <- 4
probCut <- seq(0,1,(1/nRanges))
xCut <- qnorm(probCut)

# Compute expectancy and confidence interval for each predictor range
expectancy <- expL <- expU <- array(0, dim = nRanges)
catLabel <- catLabelShort <- array("", dim = nRanges)
for (i in 1:nRanges) {
  expectancy[i] <- Expectancyfunc(Validity,xCut[i],xCut[i+1],CriterionCutoff,Inf)$expectancy
  expL[i] <- Expectancyfunc(ValidityCI[1],xCut[i],xCut[i+1],CriterionCutoff,Inf)$expectancy
  expU[i] <- Expectancyfunc(ValidityCI[2],xCut[i],xCut[i+1],CriterionCutoff,Inf)$expectancy
  catLabel[i] <- paste0(round(probCut[i]*100,0),"% - ",round(probCut[i+1]*100,0),"%")
  catLabelShort[i] <- paste0(round(probCut[i]*100,0),"%")
}

xAxisLabel <- "Predictor Score Range (Lower Bound)"
```

```
yAxisLabel <- "Percent High Performers"
x <- data.frame(expectancy,catLabel)

# Vertical bars
ggplot(x, aes(x=catLabelShort, y=expectancy, label=sprintf("%0.0f%%",expectancy*100))) +
  geom_bar(stat="identity", width = .7, fill = "seagreen2") +
  ggtitle(predictorLabel) +
  xlab(xAxisLabel) +
  ylab(yAxisLabel) +
  scale_y_continuous(labels = function(x) paste0(x*100, "%")) +
  geom_text(nudge_y = (.1), size = 4) +
  geom_errorbar(aes(ymin=expL, ymax=expU), width=.1) +
  theme_classic(base_size = 14)

# Horiszontal bars
xAxisLabel <- "Predictor Score Range"
ggplot(x, aes(x=catLabel, y=expectancy, label=sprintf("%0.0f%%",expectancy*100))) +
  geom_bar(stat="identity", width = .7, fill = "seagreen2") +
  ggtitle(predictorLabel) +
  xlab(xAxisLabel) +
  ylab(yAxisLabel) +
  scale_y_continuous(labels = function(x) paste0(x*100, "%")) +
  geom_text(nudge_y = (-.021 - .5 * abs(expU-expL)), size = 3) +
  geom_errorbar(aes(ymin=expL, ymax=expU), width=.2) +
  theme_classic(base_size = 14) +
  coord_flip()
# Save plot as an image file or copy to clipboard, then insert into word or powerpoint.

# Example 11 -----------------------------------------------------------

# Plot expectancy relative to other selection methods
# Validity for the unstructured interview set at .18, the uncorrected estimate McDaniel et al.
(1994)
library(Hmisc)  # provides cor.test function
library(ggplot2)  # graphics package
source("Cucina Expectancy Function.R") # R script for Expectancyfunc ()

# Compute validity
predictorVariable <- "SDis"  # name of variable in dataframe
predictorLabel <- "Self-Discipline" # label for focal predictor
criterionVariable <- "Performance"
rStats <- cor.test(dat[,predictorVariable],dat[,criterionVariable])
validity <- rStats$estimate

validity <- c(0,.18,validity) # vector of validies for
testLabel <- c("Random","Unstructured Inteview",predictorLabel)
testLabel <- factor(testLabel, levels = testLabel)
```

```
PredictorCutoff <- qnorm(.75)
CriterionCutoff <- qnorm(.75)
expectancy <- sapply(validity,FUN = Expectancyfunc,
                PredLowerCut = PredictorCutoff, PredUpperCut = Inf,
                CritLowerCut = CriterionCutoff, CritUpperCut = Inf)
expectancy <- unlist(expectancy["expectancy",])
result <- data.frame(testLabel,validity,expectancy)

xAxisLabel <- "Predictor"
yAxisLabel <- "Percent High Performers"
ggplot(result, aes(x=testLabel, y=expectancy, label=sprintf("%0.0f%%",expectancy*100))) +
         geom_bar(stat="identity", width = .5, fill = 'seagreen3') +
         xlab(xAxisLabel) +
         ylab(yAxisLabel) +
         scale_y_continuous(labels = function(x) paste0(x*100, "%")) +
         geom_text(nudge_y = -.1) +
         theme_classic(base_size = 14)
# Save plot as an image file or copy to clipboard, then insert into word or powerpoint.

# Example 12 -----------------------------------------------------------

# Plot expectancy for hierarchical regression
library(Hmisc)  # provides cor.test function
library(ggplot2)  # graphics package
source("Cucina Expectancy Function.R") # R script for Expectancyfunc ()

# Specify multipel R from series of regression models
validity <- c(0.40, 0.62, 0.66)

# Enter a label of reach model
testLabel <- c("Degree","Degree\n+AR+RC+BC","Degree\n+AR+RC+BC\n+SDis+SUP")
testLabel <- factor(testLabel, levels = testLabel)

# Specify passing rate and success rate
PassRate <- .25
SuccessRate <- .25
PredictorCutoff <- qnorm(1-PassRate)
CriterionCutoff <- qnorm(1-SuccessRate)
expectancy <- sapply(validity,FUN = Expectancyfunc,
                PredLowerCut = PredictorCutoff, PredUpperCut = Inf,
                CritLowerCut = CriterionCutoff, CritUpperCut = Inf)
expectancy <- unlist(expectancy["expectancy",])
result <- data.frame(testLabel,validity,expectancy)

xAxisLabel <- "Predictor Composite"
yAxisLabel <- "Proportion High Performers"
```

```
ggplot(result, aes(x=testLabel, y=expectancy, label=sprintf("%0.0f%%",expectancy*100))) +
	geom_bar(stat="identity", width = .5, fill = 'seagreen3') +
	xlab(xAxisLabel) +
	ylab(yAxisLabel) +
	scale_y_continuous(labels = function(x) paste0(x*100, "%")) +
	geom_text(nudge_y = -.1) +
	theme_classic(base_size = 14)
# Save plot as an image file or copy to clipboard, then insert into word or powerpoint.
```

Appendix C

R function for computing expectancy tables. Adapted from Cucina, Berger & Busciglio (2017). Run the script below in order to initialize the Expectancyfunc() function. Additional instructions for using the function are provided in the chapter.

```
# load library for computing normal integrals
library(mvtnorm)


# Expectancy function ----------------------------------------------------


Expectancyfunc <- function (Validity,
                    PredLowerCut, PredUpperCut,
                    CritLowerCut, CritUpperCut)
{
# This creates a new function in R called Expectancyfunc. The function takes the criterion-
related
# validity coefficient, the lower and upper cutoffs for the predictor score, and the lower and
upper
# cutoffs for the criterion score as inputs. To represent positive or negative ???, "Inf" or "-Inf"
can
# be used, respectively.
# library(mvtnorm) Before proceeding, the mvtnorm library must be downloaded and installed.
This command line
# tells R that the mvtnorm library is being used.

#  n <- 1000
#  A dataset must be created before R can be run to conduct the analyses. This command tells R
to
#  create a dataset with 1,000 cases. The value n represents the number of cases and the symbol
<-
#    indicates that n should be set equal to 1,000.
  mean <- c(0, 0)
#  In this line, the means for the two variables (which equal 0 when a standardized solution is
#  used) are provided. Note that the values are presented parenthetically, separated by a comma,
#  and preceded by the letter c. This syntax stores the means as a vector in R.
  lower <- c(PredLowerCut, CritLowerCut)
#  This line assigns the lower z-score cutoffs for the predictor and the criterion to a vector.
  upper <- c(PredUpperCut, CritUpperCut)
#  This line assigns the upper z-score cutoffs for the predictor and the criterion to a vector.
  corr <- diag(2)
#  This line creates a 2-by-2 matrix with diagonal values of 1 and stores the matrix in the
variable
#  corr.
  corr[lower.tri(corr)] <- Validity
  corr[upper.tri(corr)] <- Validity
#  In these two steps, the correlation between the two variables provided by the user is stored
into
#  the upper and lower triangles of the 2-by-2 correlation matrix.
  jtprob <- pmvnorm(lower, upper, mean, corr,
                 algorithm = Miwa(steps = 128))
  # Here the pmvnorm command in the mvtnorm package is run; this is the command that is
used
  # for computing the volume under multivariate-normal distributions. As inputs, pmvnorm
takes
  # the upper and lower z-score cutoffs (which are vectors), the vector of means (which is set to
0),
  # the correlation matrix, and the algorithm that is to be used. The algorithm statement specifies
  # that the Miwa et al. (2003) algorithm should be used. The term "(steps = 128)" informs R
that
  # 128 grid points should be used. The output for this procedure is the joint probability between
  # the predictor and the criterion - the volume under the bivariate-normal distribution between
the
  # lower and upper cutoffs. This probability is saved in the variable jtprob.
  #jtprobOutput <- paste("Joint Probability: ", jtprob, sep="")
  # This line creates a new string variable containing the value of jtprob along with a label. The
  # term "sep=""" indicates that there are no text separating the expectancy value and the %
symbol.
  #print(jtprobOutput)
  #The previous steps saved the volume of the bivariate-normal distribution and added a label;
this
  #  step prints that value, with the label, to the screen.

  #  Computing the expectancy
  xprob <- pnorm(PredUpperCut,
              mean=0, sd=1)-pnorm(PredLowerCut, mean=0, sd=1)
  # To compute the expectancy, we must obtain the proportion of cases that have a predictor
value
  # within the lower and upper cutoffs for the predictor. This is accomplished by computing the
  # area under the univariate-normal distribution, which is the proportion of cases having
predictor
  # values within the upper and lower cutoffs. The pnorm command in R is used to compute this
  # area and it takes the upper or lower predictor cutoff, mean (which is set to 0), and standard
deviation
  # (which is set to 1) as inputs. The proportion of cases that fall within the upper and lower
  # cutoffs is obtained by subtracting the proportion of cases falling between the lower cutoff
and
  # -??? from the proportion of cases falling between the upper cutoff and -???. This value is
stored to
  # a new variable, xprob.
  #xprobOutput <- paste("Predictor Probability: ", xprob, sep="")
```

```
# This line creates a new string variable containing the value of xprob along with a label.
  #print(xprobOutput)
  # This command prints the value xprob to the screen along with a label.

  expectancy <- jtprob[1]/xprob[1]
  #expectancy <- paste(round(100*jtprob/xprob,1), "%", sep="")
  # The expectancy is computed by dividing the joint probability by the predictor probability.
The
  # expectancy is converted to a percentage using the syntax "100*." In addition, this value is
  # rounded to one decimal place, using the syntax "round(.., 1)." Next, a percentage symbol is
  # added using the "paste" command, which pastes the expectancy value and the % symbol
(shown
  # in the syntax using "%") together into a string variable named "expectancy."
  #print(expectancy)

  tmp <- list(jtprob = jtprob[[1]],xprob = xprob[[1]],expectancy = expectancy[[1]])
  #print(tmp)
  return(tmp)
}
```