AFRL-RH-WP-TR-2021-0003

Air Force Personnel Center Best Practices Guide

Selection and Classification Model Development

Robert E. Ployhart

Personnel Decisions Research Institutes, LLC

January 2021

Interim Report

DISTRIBUTION STATEMENT 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-2021-0003 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. WILLIAMS
Core Research Area Lead
Performance Optimization Branch
Airman Biosciences Division

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE (DD-MM-YY)	2. REPORT TYPE	3. DATES COVERED (From - To)
19-01-21	Interim	14-03-19 – 13-12-20

4. TITLE AND SUBTITLE

Air Force Personnel Center Best Practices Guide: Selection and Classification Model Development

5a. CONTRACT NUMBER

FA8650-14-D-6500, Task Order 0007

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

62202F

6. AUTHOR(S)

Robert E, Ployhart

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/RHCC

11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)

AFRL-RH-WP-TR-2021-0003

12. DISTRIBUTION/AVAILABILITY STATEMENT

Distribution Statement A: Approved for public release

13. SUPPLEMENTARY NOTES

Report contains color. AFRL-2021-0260, cleared on 3 February 2021

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 (AFPC/DSYX). The body of the report provides recommendations and best practices in selection and classification for DSYX.

15. SUBJECT TERMS

Selection, classification, testing, Air Force, Personnel Testing, Statistical Modeling

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	86	19b. TELEPHONE NUMBER (Include Area Code) N/A

TABLE OF CONTENTS

FOREWORD iv
EXECUTIVE SUMMARY v
Introduction to the Air Force Personnel Center Strategic Research and Assessment Branch (AFPC/DSYX) vi
Background/History vi
1.0 SELECTION AND CLASSIFICATION MODEL DEVELOPMENT 1
1.1 . Introduction 1
1.1.1. Purpose of Chapter 1
1.1.2. Brief Summary of Existing Air Force Practices and Challenges. 1
1.1.3. Scope 1
1.1.4. Definitions 2
1.2 . Basic Concepts in Selection and Classification Models 3
1.3 . Selection and Classification Model Development 6
1.3.1. Overview 6
1.3.2. Needs Assessment 7
1.3.3. Job Analysis 8
1.3.4. Criterion Development 20
1.3.5. Predictor Development 22
1.4 . Techniques for Establishing Evidence of Predictive Relationships 25
1.4.1. Scores and Distributions 25
1.4.2. Estimates of Relationships 26
1.4.3. The Necessary Role of Judgment 32
1.5 . Artifacts and Conditions Affecting Predictive Relationships 34
1.5.1. Scale Coarseness 34
1.5.2. Nonnormality 34
1.5.3. Nonlinearity 35
1.5.4. Unreliability 35
1.5.5. Range Restriction 36
1.5.6. Applications and Corrections 38
1.6 . Combining Predictor Information for Selection and Classification 39
1.6.1. Combining KSAO Predictor Scores 40
1.6.2. Types of Selection and Classification Systems 45
1.6.3. Additional Practical Considerations 46
1.7 . Subgroup Differences and Adverse Impact 46
1.7.1. Overview of Diversity and Validity 46
1.7.2. Nature and Consequences of Subgroup Differences 47
1.7.3. Adverse Impact 48

1.7.4. Differential Prediction 49
1.7.5. Ways to Balance Diversity and Validity 50
1.8 . Generalizing from Experimental to Operational Use 56
1.8.1. Cross-Validity and Cross-Validation 56
1.8.2. Development of Local Norms 57
1.8.3. Retesting 57
1.9 . Saving Data and Reporting Results 57
1.10 Future Issues 58
1.10.1. Big Data 58
1.10.2. New Analytics 58
1.10.3. Big Data Recommendations 59
2.0 REFERENCES 61
APPENDIX SAMPLE CODE 67
LIST OF SYMBOLS, ABBREVIATIONS, AND ACRONYMS 69

List of Figures

Figure 1. Visual Illustration of the Difference between Constructs and Scores Based on Assessments 4
Figure 2. A Framework for Understanding Validity in Selection. Adapted from Binning and Barrett (1989) and Guion (2011) 5
Figure 3. Selection and Classification Model 7
Figure 4. The Effects of Cut Scores on Classification Decisions and Statistical Power 42

List of Tables

Table 1. Key Types of Validity Evidence for Selection 6
Table 2. Steps for Conducting a Job Analysis 8
Table 3. Sample Critical Task x Selection KSAO Matrix 11
Table 4. The "How" (Methodology) and "Who" (Sampling) of Major Job Approaches 13
Table 5. Potential Sources of Job Analysis Inaccuracy 18
Table 6. Examples of Assessment Methods and Their Associated KSAO Constructs 23
Table 7. Strengths, Weaknesses, and Potential Trade-Offs for Different KSAO Assessments 24
Table 8. Correlation Types for Nominal (Categorical) or Interval (Continuous) Relationships 26
Table 9. A Typology of Range Restriction Models (Sackett & Yang, 2000) 38
Table 10. Suggestions for Reducing Subgroup Differences and Increasing Validity (adapted from Ployhart & Holtz, 2008) 51

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 chapter describing AFPC/DSYX and the background of their research to provide context for the series. This report addresses the development of selection and classification models, including supporting topics such as job analysis and criterion development.

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 the Air Force Personnel Center Strategic Research and Assessment Branch (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 AFPC/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 AFPC/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 provides recommendations and best practices in selection and classification for AFPC/DSYX. The recommendations are based on over a century of scientific research and practice, both within the United States Air Force (USAF) and in the scientific literature more generally.

Selecting the right talent and classifying each person to the specialty and occupation that best fit their talents is vital for effective individual and organizational performance. Selection and classification are the first step in the management of talent. Consequently, every downstream activity (training, development, succession planning) benefits from more rigorous selection and classification. Performance, learning, development, retention, and satisfaction, are all improved by effective selection and classification.

We begin with a brief summary of existing Air Force practices and challenges, introducing key definitions and basic concepts. This leads into a discussion of model development including job analysis as a foundation and how to select or generate predictor and criterion measures. From here, the report describes techniques for establishing evidence of predictive relationships, including methods for handling artifacts and conditions affecting statistical estimates of those relationships, and methodologies for combining predictor scores.

Then, the report turns to a discussion of the different types of selection and classification systems, breaking them down into five broad approaches, identifying key characteristics and utility of each. This is followed by additional practical considerations including subgroup differences and adverse impact. Next we review strategies for generalizing from experimental to operational use of the selection and classification models, and we conclude with the identification of future trends to monitor going forward.

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 AFPC/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 AFPC/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 AFPC/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 AFPC/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 AFPC/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 AFPC/DSYX branch is now embedded within the AFPC Directorate of Staff. As previously mentioned, while no longer supported by a multitude of scientists and psychologists, AFPC/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 AFPC/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 AFPC/DSYX has the responsibility of developing and validating the tests within the personnel testing program, the operational responsibility of military testing resides with AFPC/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

AFPC/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 and Enlisted Classification Directories (AFOCD & 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 AFPC/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, Remotely-Piloted Aircraft (RPA) pilot combat system operators, air battle manager, or technical. AFOQT 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 Air Force Specialty Code (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

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

AFSCs. The second version is 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 USAF, 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 Pilot Candidate Selection Method (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 ATST 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 MTT 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 AFPC/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 AFPC/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 – AFPC/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 – AFPC/DSYX personnel have enlisted- and officer-level vocational interest inventories. The tools developed by AFPC/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 – AFPC/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 – AFPC/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 – AFPC/DSYX staff members have extensive experience in ethical behavior, integrity, and counterproductive behavior. AFPC/DSYX has developed integrity tests and valid tests designed to detect the propensity to engage in counterproductive behavior.

- Realistic Job Preview Creation – AFPC/DSYX staff members have extensive expertise in developing realistic job preview videos based on subject matter expert (SME) input video-based interviews.
- Leadership – AFPC.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 AFPC/DSYX

Increased Effort to have AFPC/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 the 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

AFPC/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 AFPC/DSYX to the operational entities after successful development of tests and selection/classification tools.

Repeatable and Scalable Processes

AFPC/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 SELECTION AND CLASSIFICATION MODEL DEVELOPMENT

1.1 Introduction

1.1.1. Purpose of Chapter

Selecting the right talent and classifying each person to the specialty and occupation that best fit their talents is vital for effective individual and organizational performance. Selection and classification are the first step in the management of talent. Consequently, every downstream activity (training, development, succession planning) benefits from more rigorous selection and classification. Performance, learning, development, retention, and satisfaction are improved by effective selection and classification.

This report provides recommendations and best practices in selection and classification for the AFPC/DSYX. The recommendations are based on over a century of scientific research and practice, both within the United States Air Force (USAF) and in the scientific literature more generally.

1.1.2. Brief Summary of Existing Air Force Practices and Challenges.

The Air Force Personnel Center is responsible for managing a workforce of over 600,000 active duty members, civilians, and reserve personnel (https://www.af.mil/About-Us/Fact-Sheets/Display/Article/104554/air-force-personnel-center/). In terms of selection and classification, standardized aptitude assessments such as the Armed Services Vocational Aptitude Battery (ASVAB) and the Air Force Officer Qualifying Test (AFOQT) are administered to candidates, followed by a battery of physical and mental assessments. Scores on these assessments are used to determine which jobs provide the best fit. Selection and classification models have been developed that maximize accuracy and diversity. However, quickly evolving technologies, assessment methods, and analytic methods make it challenging to stay current and know whether newer approaches offer value over traditional methods of known effectiveness.

1.1.3. Scope

The scope of this report is on the process of selection and classification. There are many other activities and practices that inform the process of selection and classification. These include psychometrics and measurement development, criterion (performance measurement) development, legal and ethical issues, statistics, and so on. These topics are briefly recognized, but it is beyond the scope of this report to review them in depth. It is assumed that readers are already familiar with the basics of validity, statistics, psychometrics, and related topics.

This report is intended to provide recommendations based on the best scientific evidence, professional practice guidelines, and legal and ethical principles. It is intended to provide a lasting set of recommendations that apply to USAF selection and classification in any situation. This report is not specific to existing USAF practices and procedures. This is intentional. Practices and procedures may change with time and technology, but the recommendations provided in this report are expected to endure long into the future.

Note this report is *not* intended to discuss or review medical assessments. Medical assessments fall outside the scope of employment testing and assessment.

1.1.4. Definitions

Personnel selection and classification are related but distinct processes. *Personnel selection (selection)* is the process of (a) defining the talent needed to perform effectively on critical job tasks, (b) identifying which candidates have the job-relevant talent, and (c) making selection decisions based on job-relevant information (Guion, 2011; Ployhart, Schneider, & Schmitt, 2006). For example, in a military setting, this could involve selecting candidates into the Air Force who have minimal qualifications for service. This would not necessarily mean they have an aptitude for any Air Force career, however.

Classification is the process of determining which types of talent are needed for different jobs and occupational specialties. That is, out of a number of different jobs and occupations, which types of talent are most needed for each of the different jobs? Classification is thus a sorting procedure whereby a candidate is linked to the types of jobs that best match their capabilities (Sellman, Russell, & Strickland, 2017). In the example above, we would essentially be sorting the candidates based on their respective aptitudes and how they match particular jobs within the Air Force.

Knowledge, skills, abilities, and other characteristics (*KSAOs*) summarize the types of talent needed to perform a job. According to the Occupational Information Network (O*NET; https://www.onetcenter.org/content.html):

- *Knowledge* is "Organized sets of principles and facts applying in general domains" of information. Examples include knowledge of a job, an organization's procedures, equipment, and so on.
- *Skills* are "Developed capacities that facilitate learning or the more rapid acquisition of knowledge." Examples include reading comprehension, writing, critical thinking, and learning strategies.
- *Abilities* are "Enduring attributes of the individual that influence performance." Examples include cognitive abilities, psychomotor abilities, physical abilities, and sensory abilities.
- *Other Characteristics* represent the variety of individual differences that do not fall neatly into knowledge, skills, or abilities. These other characteristics may include:
  - Personality (known as "Work Styles" in O*NET) refers to "…a habitual way of thinking or doing in a variety of situations." (Guion, 2011, p. 105). Examples include openness to experience, conscientiousness, extraversion, agreeableness, and emotional stability.
  - Work Values are "Global aspects of work composed of specific needs that are important to a person's satisfaction." Examples include achievement, independence, and recognition.
  - Occupational Interests are "Preferences for work environments." Examples include realistic, investigative, artistic, social, enterprising, and conventional interests.

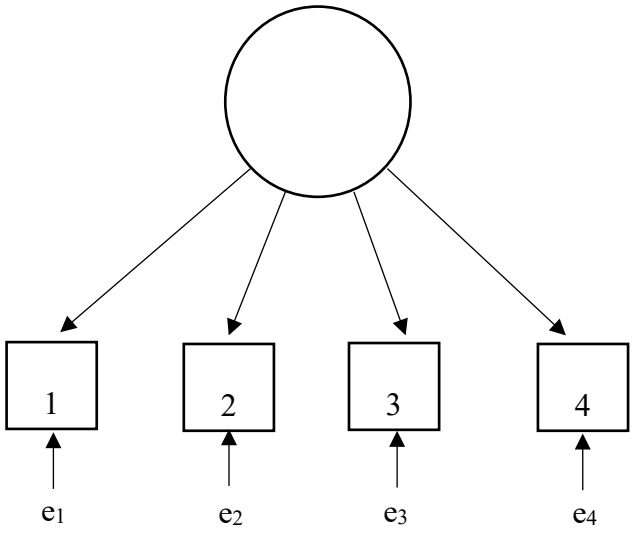
Note that historically, the science and practice of selection and classification focused on knowledge, skills, and abilities (Ployhart, Schmitt, & Tippins, 2017). It was not until the early 1990s that research began to support the inclusion of personality, work values, and occupational interests into the selection and classification model (Sackett, Lievens, Van Iddekinge, & Kuncel, 2017). The fact that so many individual difference constructs are collapsed into the "Other" category should not be taken as a sign they are unimportant; it simply reflects the conventions of selection research and practice.

Note that key outcomes of selection and classification are to enhance performance, retention, and satisfaction. For purposes of this report:

- *Performance* is usually defined in terms of job performance and is "… the things people actually do, the actions they take, that contribute to the organization's goals." (Campbell & Wiernik, 2015, p. 48). Performance is different from KSAOs in that it is the behaviors and actions in which an individual engages, and is a manifestation of the KSAOS that individual brings to the role. Performance is also not the same as the outcomes or results of performance (e.g., accidents, bookkeeping accuracy, or productivity). They are related but not the same in that many external factors unrelated to performance can impact outcomes. (Campbell & Wiernik, 2015). There are different types of job performance that include task performance, citizenship performance, adaptability, and counterproductive work behaviors (a negative type of performance).
- *Criteria* are measures of performance.
- *Retention* refers to the person staying gainfully employed within the job and/or organization. Turnover occurs when a person leaves the organization, either voluntarily or involuntarily.
- *Satisfaction* is "the emotional state associated with the self-evaluation of work." (Locke, 1976).

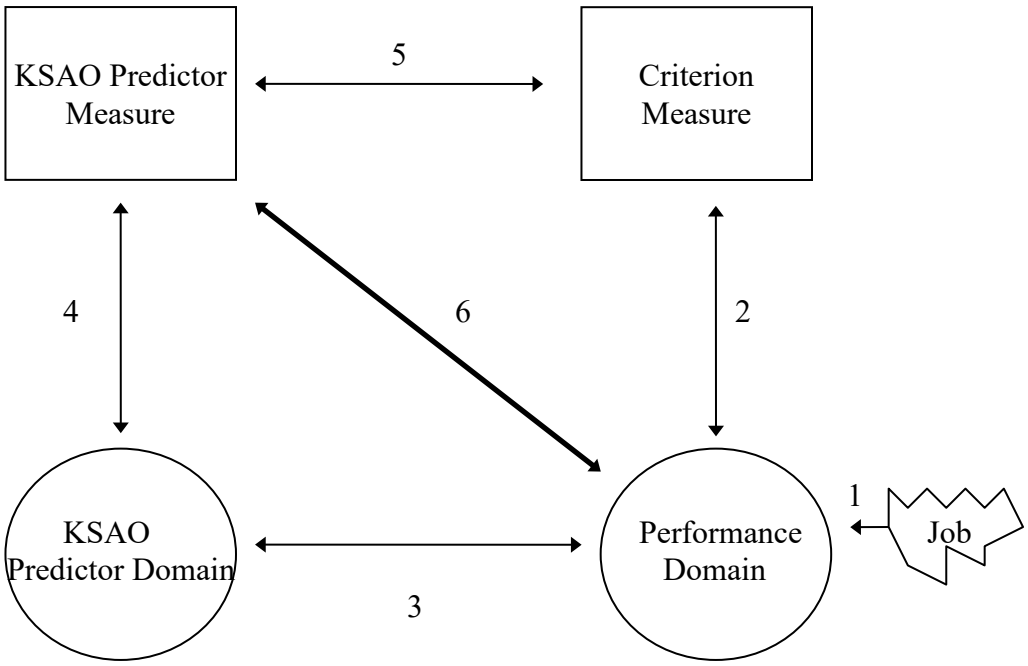
1.2 Basic Concepts in Selection and Classification Models

The most basic purpose of selection and classification is to use job-relevant information to make a prediction about a person's future performance (Ployhart, Weekley, & Dalzell, 2018). The information used to make such predictions is based on scores provided by measures, tests, assessments, or related indicators of latent KSAO constructs, as illustrated in Figure 1. For example, a cognitive test, an interview, or a flight simulation are all examples of assessments of latent constructs that provide scores that can be used to make selection and classification decisions. It is important to understand these elements because they comprise the building blocks to any selection and classification system. Figure 2 provides an overview of these assessment elements that are discussed in more detail below.



The circle represents the latent (unobservable) construct, arrows represent hypothesized causality, the boxes represent observed variables, the four items that produce scores, and the symbol "e" refers to error. Thus, this figure illustrates how each item score is a function of two sources of variance: the true score (from the circle) and error.

Figure 1. Visual Illustration of the Difference between Constructs and Scores Based on Assessments



Forms of Validity Evidence. Boxes indicate manifest (measured) variables, circles represent latent (unmeasurable) constructs, and the "job" shape represents how the nature of the job is fundamental to defining the performance domain and thus various forms of validity evidence.
Arrow 1: Content validity (performance domain is linked to critical job tasks via job analysis)
Arrow 2: Construct validity (of criterion scores)
Arrow 3: Hypothesized relationship between latent KSAO predictor domain and latent performance domain (based on job analysis)
Arrow 4: Construct validity (of KSAO predictor scores)
Arrow 5: Criterion-related validity (correlation between KSAO predictor scores and criterion scores)
Arrow 6: Inference that scores on KSAO predictor are associated with true changes in performance.

Figure 2. A Framework for Understanding Validity in Selection. Adapted from Binning and Barrett (1989) and Guion (2011)

An assessment produces scores, and it is the scores that are used to make decisions and form inferences. The distinction between scores and assessments is critical because scores, not assessments, are used to form inferences about validity (American Psychological Association (APA) Standards, 2014; Messick, 1995; Murphy, 2012; Schmitt, Arnold, & Nieminen, 2017). Validity is defined as "…the degree to which evidence and theory support the interpretation of test scores for proposed uses of the test" (APA Standards, 2014, p. 11). Note that a score provided by an assessment may have many potential uses. AFOQT scores, for example, may be used to make classification decisions (e.g., into which occupation should this person be placed),

predict retention and satisfaction in the job, and determine training needs (e.g., what types of training are needed to prepare this person for the job).

Validity is an inference about whether scores are appropriate or informative for some specific purpose. *Validation* involves collecting evidence to support an inference (in selection, this inference is arrow 6 in Figure 2). However, there are multiple kinds of validity, and each type helps support an overall inference of validity. Figure 2 and Table 1 provide an overview of these different forms. Note that Arrow 1 in Figure 2 is used to represent the fact that the requirements of the job (identified via job analysis) define the performance domain. It is beyond the scope of this report to go into them in greater detail (see APA Standards, 2014; Messick, 1995; Murphy, 2009; for more details).

Table 1. Key Types of Validity Evidence for Selection

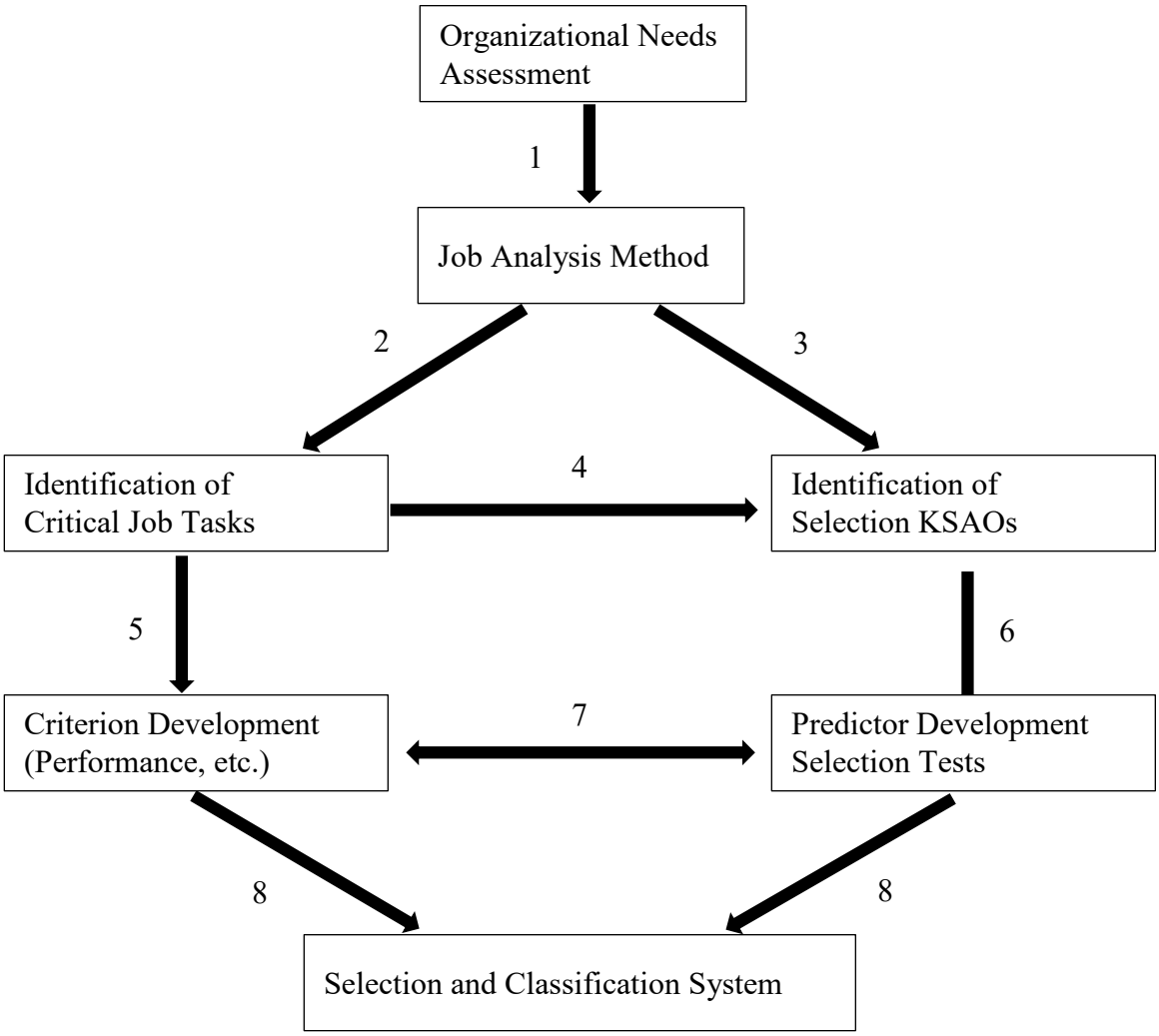
Criterion-related: Relationship between KSAO predictor scores and criterion scores. Estimated using correlation or regression.
• **Concurrent:** Correlation between predictor scores and criterion scores when both variables are collected on job incumbents. More efficient than predictive validity, but does not allow for as strong an inference of validity.
• **Predictive:** Correlation between predictor scores and criterion scores, but predictor variable is collected with job applicants and criterion variable is collected after some of applicants are hired and become job incumbents. Less efficient than concurrent validation because of the time and tracking of incumbents required, but allows for stronger inferences of validity.
Content: Expert judgment about the conceptual overlap between predictors and criteria. Based on job-analysis.
Construct: Evidence supporting an inference between scores on an assessment and the KSOA/performance construct the scores are supposed to represent.
• **Convergent**: Scores are positively associated with scores representing similar attributes but measured in different ways.
• **Discriminant**: Scores are weakly associated with scores representing different attributes.

Note: Adapted from Ployhart et al., 2018

1.3 Selection and Classification Model Development

1.3.1. Overview

The development of selection and classification models follows a specific sequence (Guion, 2011; Schmitt & Chan, 1998). This process involves a systematic and reasonably comprehensive approach to ensure that selection and classification decisions are based only on the KSAOs necessary for critical jobs tasks (Society for Industrial and Organizational Psychology SIOP Principles, 2018). In this manner, the entire process ensures that selection and classification are job-related. Following this process will lead to selection and classification decisions that are more predictive and less biased. This process is shown in Figure 3 and detailed below.



Note: Adapted from Ployhart et al., 2006; Schmitt & Chan, 1998

Figure 3. Selection and Classification Model

1.3.2. Needs Assessment

Needs assessment provides the organizational strategic overview of the entire selection and classification system. As such, needs assessment seeks to forecast current and future talent needs, staffing levels needed in different occupations, changes to the nature of work due to technology, societal and political forces that may influence recruitment and workforce planning, and all other relevant economic and workforce issues (e.g., strategic vision for agile total force human capital management). Needs assessment helps determine which occupations are strategic and which are operational, as a means to better utilize and deploy key resources. Needs assessment will also help identify when it is time to re-evaluate the selection and classification model, or to redo a job analysis. Thus, selection and classification models are embedded within the broader workforce

planning strategy (Brannick, Pearlman, & Sanchez, 2017; SIOP Principles, 2018). This fact is highlighted in Figure 3 (arrow 1).

1.3.3. Job Analysis

Job analysis is the foundation of any selection and classification model. Job analysis (or work analysis) is defined as the "…systematic process for gathering, documenting, and analyzing information about (a) the content of the work performed by people in organizations (e.g., tasks, responsibilities, or work outputs), (b) the worker attributes related to its performance (often referred to as KSAOs), or (c) the context in which work is performed (including physical and psychological conditions in the immediate work environment, and the broader organizational and external environment)" (Brannick et al., 2017, p. 134).

The purpose of job analysis is to identify the critical KSAOs needed to perform the critical tasks of a job. A thorough job analysis will capture the important work-related aspects of a job, including tasks and activities, tools and equipment, the broader context, and thus the KSAOs required to perform effectively. A poorly done or incomplete job analysis will result in less accurate selection and potential bias in selection and classification decisions. There are five broad steps to conducting a job analysis (these are shown in Table 2).

Table 2. Steps for Conducting a Job Analysis

Step	Purpose
1. Comprehensively Identify Tasks	• Create comprehensive list of tasks using a variety of methods.
2. Identify Critical Tasks	• Rate the criticality, importance, frequency, time-spent, and/or consequences of mistakes, etc., on each task (done by subject matter experts). • Identify the smaller subset of truly critical tasks. • Group critical tasks into approximately 5–15 task clusters.
3. Use Critical Tasks to Identify KSAOs	• Create comprehensive list of KSAOs linked to tasks: abilities, personality, knowledge, skills, etc. • Each type of KSAO is linked to a critical task.
4. Identify the Selection KSAOs	• Have subject matter experts rate each KSAO in terms of importance for performing the job and whether it is needed at the start of the job. • Reduce the total list of KSAOs into a smaller set containing only critical KSAOs (called selection KSAOs).
5. Develop Task × KSAO Matrix	• Show the linkages between each task and each Selection KSAO. • This step is usually conducted by HR personnel and reviewed by subject matter experts.

Note: Adapted from Ployhart et al., 2018.

Step 1. The process starts by comprehensively understanding the nature of the job. This is done through reviewing existing documentation about the job, including technical and procedures manuals, existing job descriptions and specifications, and any related information (such as O*NET). It is also important to observe incumbents performing the work, and interviewing job incumbents, supervisors, and all other subject matter experts (SMEs) to understand how the work is performed and the context within which it is performed. Note that SMEs may provide different perspectives on the work being performed. Job incumbents are best at describing how the work is actually performed. Supervisors often focus more on describing how the work is supposed to be performed. Both perspectives can be appropriate; and different ways of doing the work may produce similar or even identical results. Hence, it is best to capture as many perspectives as possible.

Regardless of who is participating, start by having SMEs list as many tasks as they feel describe their job. When they can no longer identify any tasks, ask them to identify tasks that an excellent employee performs and the tasks that a poor employee performs, to see if any new tasks are identified. One may then introduce any written documentation surrounding tasks, and ask the SMEs to review it to see if any tasks should be added or eliminated.

The conclusion of Step 1 is a comprehensive description of all job tasks. These tasks may number in the hundreds. Each task should be captured in a single sentence known as a task statement. The task statement should start with a verb that describes a concrete action, the context within which the task is performed, and any additional description (Barrick et al., 2017; Guion, 2011; Ployhart et al., 2006). An example may include: "Monitors radar equipment in command center to ensure safe operation of airspace." There are many approaches to writing task statements; different approaches can be found in several sources, including Guion (2011) and Gatewood, Feild, and Barrick (2011). The Office of Personnel Management (OPM) provides considerable guidance on how to conduct job analysis (OPM, 2019; see also https://www.opm.gov/policy-data-oversight/assessment-and-selection/job-analysis/).

Step 2. A comprehensive listing of task statements is vital to ensuring nothing is missed in developing the selection and classification model. However, the number of tasks is often unwieldy and many of the tasks are not necessarily important. The goal of Step 2 is to winnow the set of tasks to a more manageable number, by identifying the tasks that are critical, important, have important consequences, and/or are performed frequently. The goal is to reduce the tasks to a grouping of around 5-15 task clusters (groups of similar tasks), with each task being important.

Practical constraints dictate how the critical tasks are identified (SIOP Principles, 2018). For example, if there is a sufficient number of job incumbents, it may be most effective to administer a task survey and have them rate the criticality, frequency, and importance of each task. Other times, it may be necessary to work with small groups to have them evaluate the tasks. Regardless of how it is done, it is important that the sample is representative of those in the job, comprised of SMEs, and quantified so that it is possible to apply analytics or simply to make more fine-grained distinctions among tasks. Sample rating scales are shown below (see Harvey, 1991; Guion, 2011; Ployhart et al., 2018; Schmitt & Chan, 1998; for other examples):

- Difficulty: How difficult is this task to perform (1 = easy, 5 = extremely difficult)

- Criticality: How critical is this task (1 = not critical, 5 = extremely critical)
- Frequency: How frequently do you perform this task (0 = never, 5 = hourly)
- Time Spent: How much time do you spend performing this task (1 = very infrequent, 5 = very frequent)
- Importance: How important is this task (1 = not important, 5 = very important)
- Consequence of Error: What are the consequences of incorrect performance on this task (1 = errors are not important, 5 = errors are extremely important)

A common approach used to provide an overall index of task importance is (Schmitt & Chan, 1998, p. 47):

Task Importance = Difficulty x Criticality + Time Spent

One may calculate task importance as noted above, cluster the tasks according to common elements, and then rank order the specific tasks within each cluster. This is why quantifying the task statements is so helpful; it becomes possible to employ analytics to identify the critical tasks according to some threshold. The conclusion of Step 2 is the identification of a smaller, manageable cluster of critical tasks.

Step 3. Once the critical tasks are established, it becomes possible to identify the job-relevant KSAOs. One starts by looking at each task and having SMEs (job incumbents, supervisors, and psychologists) identify the potential KSAOs that are needed to perform that task. Usually, performance on a given task will be influenced by several KSAOs. The purpose of this step is simply to identify as many KSAOs as relevant to ensure full coverage of the KSAO domain.

Start by having the SMEs list as many KSAOs as they feel are relevant for each task. They should do this using their own vocabulary and opinions. When they can no longer identify the KSAOs, ask them to identify the KSAOs of highly effective employees, and then ask them to identify the KSAOs of ineffective employees, to see if any new KSAOs are identified. It is often difficult for incumbents and supervisors to think in terms of KSAOs because they are not trained or experts in individual difference constructs. Therefore, after they provide their own opinion, it's often helpful to then provide the SMEs with a list of potential KSAOs defined based on existing documentation and the scientific literature. Ask the SMEs to review this list to identify any KSAOs that may be missing, and link these KSAOs to tasks. The conclusion of this step is a comprehensive list of potential KSAOs needed for the job.

Step 4. This step seeks to winnow the list of KSAOs to those that are truly critical for preforming the job—that is, identify the "selection KSAOs" or those that will be used as a basis for selection and classification (Schmitt & Chan, 1998). Similar to Step 2, it is best to quantify the identification of selection KSAOs whenever possible, so that fine distinctions and analytics can be used. Ratings can be used such as those noted below (see Schmitt & Chan, 1998):

- Importance: How important is this KSAO for performing this task (1 = completely unimportant, 5 = extremely important)
- To what extent does this KSAO distinguish a superior worker from an average worker (1 = very little or no extent; 5 = extremely or great extent)

Sometimes additional insight into the KSAOs is desired. For example, it may be valuable to identify the level of competence required on each KSAO. This is oftentimes helpful for creating classification models, where different jobs can be contrasted in terms of their KSAO requirements, even for the same KSAO. For example, the vision requirement may be greater for a pilot than an office worker. Sample rating scales for this purpose are shown below:

- Proficiency: How proficient must a minimally competent employee be in using this KSAO (1 = low; 5 = mastery)
- Competence: What degree of competence on this KSAO is required for this task (1 = novice, 2 = advanced beginner, 3 = competent, 4 = proficient, 5 = expert)

The conclusion of this step is the identification of the selection KSAOs, linked to each critical task that will comprise the basis of selection and classification.

Step 5. The final step involves creating a critical task x Selection KSAO matrix. Such a matrix does more than make explicit the linkages between each critical task and each selection KSAO. It can also be used to identify the relative importance of each KSAO to the overall performance of the job. This information, in turn, can be used to determine test content, aid in writing items, and so on. Table 3 provides an example matrix containing three critical tasks and four Selection KSAOs.

Table 3. Sample Critical Task x Selection KSAO Matrix

Critical Tasks	**Selection KSAOs**			
	KSAO1	***KSAO2***	***KSAO3***	***KSAO4***
Task1	X	X		
Task2	X		X	X
Task3	X			X
Relative importance:	(3/7) = 43%	(1/7) = 14%	(1/7) = 14%	(2/7) = 29%

In this example, each "X" in a cell indicates a selection KSAO linked to a critical task. The relative importance of each KSAO is determined by the following equation:

KSAO relative importance = Σ "x's" in each column / Σ of cells with an "x"

For example, KSAO 1 is needed for all three critical tasks, and hence its relative importance to the total job is 43%. KSAO 4 is second most important (29%), and KSAOs 2 and 3 are each tied at 14%. Should we be developing a 100-question assessment, 43 questions will tap KSAO 1, 14 questions will tap KSAO 2, 14 questions will tap KSAO 3, and 29 questions will tap KSAO 4. The conclusion of Step 5 thus provides a blueprint for the construction of a selection and classification model.

A common challenge in performing Step 5 is for the SMEs to assign tasks to KSAOs. For example, SMEs will often indicate a task is linked to every KSAO in the list. One approach that

works reasonably well is to set rules for making the linkages shown in Table 3. For example, a rule might be "you can assign no more than three KSAOs to a task (i.e., no more than three "X's" in a row). Alternatively, after making their assignments, a rule may be "Now try to reduce the number of "X's" in a row to the smallest set possible." Such rules are only used as quality checks and to have SMEs critically evaluate their linkages. One wants the SMEs to make the linkages according to their own experience; the rules are only there to ensure they are giving each linkage full consideration.

Additional Considerations in Job Analysis. As noted, it is critical that job analysis be performed in a systematic comprehensive manner, balanced against practical realities and constraints (SIOP Principles, 2018). To ensure job analysis is performed in the most effective manner possible, make sure the following issues are considered and defensible.

Methods of collecting job analysis information. There are a variety of approaches for collecting job analysis information: interviews, focus groups, critical incident approaches, observation, surveys, and so on. Table 4 lists different data collection approaches, and their potential strengths and weaknesses. Additional information on job analysis data collection methods, including sample instruments and procedures, can be found in O*NET (https://www.onetcenter.org/) and OPM (https://www.opm.gov/policy-data-oversight/assessment-and-selection/job-analysis/).

Table 4. The "How" (Methodology) and "Who" (Sampling) of Major Job Approaches

Job Analysis Approach	Strengths	Limitations	Comments
How (methodology)			
Review of Existing Job Documentation and Content	• Provides baseline information of job. • Economical use of time and resources.	• Usually only a broad summary; not comprehensive. • Assumes everyone performs the job similarly. • Describes how job should be done, not how it is actually done.	• Provides a helpful starting point but insufficient without supplementing with additional information. • O*NET website is almost always a good place to start.
Job Observation	• Economical use of employee time and resources. • Can see how work is performed, with what tools, and in what context.	• Observing behavior may change behavior. • Observing behavior in some jobs is difficult or dangerous. • Some knowledge work is not really observable.	• Helpful to at least observe some samples of actual work behavior, if for no other reason than to provide context.
Critical Incidents	• Structured technique that identifies (a) antecedent of behavior, (b) behavior, and (c) consequence of behavior. • Can be collected by interviewing employees or observing behavior. • Provides detailed information about tasks.	• Same potential limitations as job observation. • Can be difficult to provide detailed information about specific tasks. • Can be expensive and time consuming to implement.	• Very useful technique for developing simulations because the work context is a part of the observation. • Often helpful to at least observe or interview some employees using this technique.

Job Analysis Approach	Strengths	Limitations	Comments
Focus Group Meetings	• Fast and efficient way to collect information. • Allows different perspectives to be considered and discussed. • Nuances between how people perform job more easily identified.	• If participants do not feel they can trust the interviewer, then they will not provide accurate information. • Can be difficult to reach consensus about tasks or types of talent. • Sometimes difficult to stimulate discussion.	• Most important issues are to (a) ensure a representative sample of employees and (b) ensure they are open and honest in their discussion. • Usually conduct sessions in groups of 5–8. • Conduct meeting with peers only; no supervisors present.
Surveys	• Allows broad access and input to job analysis. • Can conduct quantitative analysis and make precise specifications. • Can be cost-effective if administered over the Internet. • Increases access and participation. • Large amounts of data collected quickly.	• Sometimes generates low response rates. • Participants sometimes do not complete survey honestly or accurately. • Requires knowledge of survey design and analysis. • Requires large samples.	• Need to ensure the sample is representative. • Evaluate validity of scores to identify response bias, faking, etc. • Try to use surveys whenever possible because they provide the most comprehensive and inclusive view of the job.

Job Analysis Approach	Strengths	Limitations	Comments
Who (Sample)			
Job Incumbents	• You must obtain information from incumbents. Because they are the ones actually doing the work, they are the only ones who can say how it is performed. • May identify important differences between employees.	• It can be expensive and time consuming to pull employees off their jobs. • Employees can be reluctant to provide information about how they perform their jobs.	• Only job incumbents can describe how the job is actually done. • Let job incumbents describe how they do the job honestly; don't lead them to describe how it *should* be done. • You must have incumbents provide job analysis information about the job tasks and behaviors. • Different incumbents may perform job differently; these differences must be recognized as real and potentially important.
Supervisors	• Often provide a unique perspective about the nature of work. • Help identify contextual or coordination challenges about the job.	• It is expensive and time consuming to pull supervisors off their jobs. • May be hesitant to describe job as it exists, rather than what they want it to be.	• Best to describe how job should be done. • Offer a useful perspective but not critical for a job analysis.
Direct Reports	• Sometimes helpful for understanding a "bottom-up" view of the job. • Provide a different perspective, particularly if the job has extensive management or leadership elements.	• It can be expensive and time consuming to pull employees off their jobs. • May be hesitant to describe job as it exists.	• Offer a useful perspective but not critical for a job analysis.

Adapted from Ployhart et al., 2018, Table 2.3, pages 86-90.

- Perceptions in job analysis. Because so much of job analysis is based on the perception and opinion of SMEs, it can sometimes be difficult to determine whether any differences across SMEs are real or perceived. If two SMEs differ in the manner in which they perform the job, the questions become first determining whether the differences matter (in terms of behavior or performance), and whether the differences are due to race, age, sex, and so on. Morgeson and Campion (1997) discuss these potential sources of inaccuracy in great detail. Table 5 provides an excerpt from their articles.
- Important information to include in a job analysis report. Both the SIOP Principles (2018) and the OPM (https://www.opm.gov/policy-data-oversight/assessment-and-selection/job-analysis/) provide guidance about the information that should be included in every job analysis report.
- Note that job analysis is different from competency modeling. Job analysis is intended to provide a comprehensive description of task and KSAO information specific to a job. Job analysis is thus job specific. Competency modeling is broader and serves a different purpose (namely, longer-term development and training). Competency modeling usually focuses on identifying a broader set of KSAOs (termed competencies) that are relevant across jobs and/or hierarchical levels in a firm. The identified competencies are thus more general (e.g., leadership) and smaller in number. Competency modeling usually does not have the degree of comprehensiveness and rigor needed to identify selection KSAOs, and in such instances is insufficient for providing a foundation for selection systems. However, if competency models are created with sufficient rigor, they may (in limited circumstances) help establish the job-relatedness of a selection system (see Campion, Fink, Ruggeberg, Carr, Phillips, & Odman, 2011, for a comprehensive discussion).

Table 5. Potential Sources of Job Analysis Inaccuracy

Table 1

Social and Cognitive Sources of Potential Inaccuracy and Their Hypothesized Effects on Job Analysis Data

Source of inaccuracy	Likely effect on job analysis data					
	Interrater reliability	Interrater agreement	Discriminability between jobs	Dimensionality of factor structures	Mean ratings	Completeness of job information
Social sources						
Social influence processes						
Conformity pressures	✓	✓				
Extremity shifts		✓		✓	✓	✓
Motivation loss			✓	✓		✓
Self-presentation processes						
Impression management					✓	
Social desirability					✓	✓
Demand effects		✓			✓	
Cognitive sources						
Limitations in information processing systems						
Information overload	✓		✓	✓		✓
Heuristics			✓	✓	✓	✓
Categorization			✓	✓		✓
Biases in information processing systems						
Carelessness	✓		✓	✓		
Extraneous information					✓	
Inadequate information	✓					✓
Order and contrast effects						✓
Halo				✓	✓	
Leniency and severity				✓	✓	
Methods effects	✓[a]			✓		

Note. Check marks indicate the likely effects the sources of inaccuracy will have on job analysis data.

[a] Refers to internal consistency reliability in this case.

Note: Excerpt from Morgeson & Campion, 1997, pages 629 and 632, respectively. Reprinted with permission from publisher.

Table 2

Job Analysis Facets and the Hypothesized Likelihood That Different Psychological Processes Will Produce Inaccuracy

	Social influence processes			Self-presentation processes			Limitations in information processing systems			Biases in information processing systems						
Facet	Conformity pressure	Extremity shifts	Motivation loss	Impression management	Social desirability	Demand effects	Information overload	Heuristics	Categorization	Carelessness	Extraneous information	Inadequate information	Order & contrast	Halo	Leniency & severity	Method effects
Job descriptor[a]																
Job-oriented	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Worker-oriented	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Analysis activity																
Generate			✓	✓	✓	✓		✓	✓		✓	✓				
Judge	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Method of data collection																
Group meeting	✓	✓	✓	✓	✓	✓					✓					
Individual interview	✓			✓	✓	✓					✓		✓			
Observation	✓			✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓
Questionnaire	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓	✓
Source of data																
Incumbent	✓	✓	✓	✓	✓	✓	✓			✓				✓	✓	✓
Supervisor		✓					✓							✓	✓	✓
Analyst								✓	✓		✓	✓	✓	✓		✓
Purpose																
Compensation		✓		✓	✓		✓				✓	✓	✓	✓	✓	✓
Selection	✓	✓	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Training	✓	✓	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Note. Check marks indicate a higher likelihood that this source of inaccuracy will result.

[a] The likelihood of inaccuracy is expected to be greater for worker-oriented than job-oriented descriptors.

Note: Excerpt from Morgeson & Campion, 1997, pages 629 and 632, respectively. Reprinted with permission from publisher.

1.3.4. Criterion Development

The next step in creating the selection and classification model is developing the measures of performance—known as *criterion development*. Criterion development must be embedded in the job analysis, as shown in Figure 3 (arrow 5). In fact, the critical tasks identified in the job analysis often become the performance dimensions used in performance management systems. The reason is because the critical tasks define what job incumbents actually do. Thus, each critical task becomes a criterion dimension that is used to define performance and assess performance. Useful criteria are relevant (i.e., job related), reliable, and discriminable (between people, and between different dimensions and levels of performance) (Ployhart et al., 2006).

Campbell and Wiernik (2015) summarize a great deal of research that has been conducted on understanding the conceptualization and measurement of job performance. First, when conceptualizing performance, it's important to realize that performance is behavior and what employees actually do. Results are the consequences of performance, and effectiveness is one's judgment of whether the results are good or bad (Smith, 1976). Campbell and Wiernik (2015) note that there are eight dimensions of job performance that, in some combination, are present for most jobs:

1. Technical performance
2. Communication
3. Initiative, persistence, and effort
4. Counterproductive work behavior
5. Supervisory, managerial, and executive leadership
6. Hierarchical management performance
7. Peer/team member leadership performance
8. Peer/team member management performance

It is important to recognize that these performance dimensions are stated in general terms that are intended to generalize across jobs and organizations. However, most specific measures of performance can usually be mapped into one of these eight types. Distinctions between management and leadership can be subtle and may overlap, but generally management involves providing guidance, support and resources, whereas leadership traditionally involves strategic, high level direction to align work with a vision or overall strategy. Further, these eight types generally load onto one superordinate factor (Viswesvaran, Schmidt, & Ones, 2005), and hence it is common to examine overall performance.

Second, when measuring performance, the difficult challenge is converting the task information into a performance measurement system that adequately captures the key elements of the performance domain. One should strive to use a measurement system that provides interval quality scores (described shortly), so that analytics can be used to make fine distinctions. There are a variety of ways of measuring performance: behaviorally anchored rating scales, behavioral observation scales, behavioral expectation scales, graphic scales, and so on. The optimal method will depend on the type of work being evaluated, the purpose of the ratings and the familiarity of raters with the performance. The common feature across these different measurement systems is that they rely on human judgments about employee behavior. These judgments are improved

considerably when the measurement system defines different levels of performance in behavioral terms. Therefore, it is usually best to use behaviorally-anchored rating scales (BARS) whenever feasible. For example, BARS will provide a number associated with each type of performance behavior on the same dimension. An example for the performance dimension bookkeeping is shown below (higher numbers indicate better performance):

5 = records entries with no errors; entries recorded on time
4 = records entries with infrequent errors; entries may be late by one day
3 = records entries with occasional errors; entries may be a few days late
2 = records entries with frequent errors; entries may be a week late
1 = records entries with consistent errors; entries may be several weeks or more late

When developing criteria, one should start by defining the performance dimension in behavioral terms. Then, with a group of SMEs, identify specific behavioral examples of each level of performance on the intended measurement scale. Have a different group of SMEs then look at a scrambled set of behaviors, and have them assign them to each number on the rating scale. Revise as necessary and continue this process until there is reasonable consensus about the behaviors associated with each number, on each performance dimension. The approach described by Smith and Kendall (1963) remains highly appropriate and should be followed whenever feasible.

When developing criteria, it is important to ensure the measures are as free from contamination and deficiency as possible. *Contamination* occurs when variance unassociated with the performance dimension influences the scores. *Deficiency* occurs when relevant (true) performance variance is not part of the measurement system. Continuing the bookkeeping example, contamination will occur if the employee's race influences the rating of bookkeepers, and deficiency will occur if accuracy is not part of the performance evaluation.

Obviously, when performance ratings are based on human judgment (e.g., supervisory or peer evaluations), there is the opportunity for contamination and deficiency to occur in the form of rater biases. There are several such biases to consider:

- Halo: the rater gives generally all positive or negative ratings across different performance dimensions for a given employee.
- Leniency: the rater gives more positive ratings to employees than other raters.
- Severity: the rater gives more harsh ratings to employees than other raters.
- Central Tendency: the rater avoids making difficult decisions by rating employees as average.

There are several steps one can take to increase the reliability and discriminability of performance behaviors:

- Train raters: Familiarize raters with the performance dimensions and criterion measures and teach them how to appropriately observe behavior.(For more on rater training techniques and their effectiveness, see Woehr & Huffcut, 1994.)
- Accountability. Ensure raters are sufficiently motivated to provide accurate ratings, are held accountable for their ratings, and have sufficient opportunities to observe behavior.

- Measurement system: Ensure the criterion measurement system is easy to understand and use, creates interval level scores, and is expressed in behavioral terms.
- Multiple ratings: Providing more ratings and/or raters increases reliability. However, the number of performance dimensions and ratings should be kept to the minimal number necessary to avoid overburdening raters.

Because of potential (or perceived) issues with subjective performance ratings, many prefer to use objective criteria. *Subjective criteria* scores are based on human judgment (as noted above); *objective criteria* scores are not created based on rater judgment. Objective criteria may include accidents, turnover rates, scrap rates, errors, days late or absent, and so on. Objective criteria are not necessarily better than subjective (ratings) criteria. First, objective criteria may also be contaminated and deficient, such as when bookkeeping errors are due to a faulty software system. Second, objective criteria may be difficult to predict when base rate events occur infrequently and hence have limited variability and nonnormal distributions. Objective and subjective criteria only correlate approximately .25 to .45, even when the performance dimensions are conceptually similar (Bommer, Johnson, Rich, Podsakoff, & MacKenzie, 1995; Rich, Bommer, MacKenzie, Podsakoff, & Johnson, 1999). Hence, it is preferable to use both types of criteria whenever possible.

Finally, it is important to recognize that criteria are the target of any selection and classification model. This means that any contamination or bias that are in the criterion measures will be reflected in the predictors used to select and place candidates. Likewise, any deficiency in the criterion measures will result in the predictor model being deficient and possibly missing important KSAO predictors. Hence, any contamination or deficiency in criteria will create contamination or deficiency in predictors. This can be observed in arrow 4 (Figure 3), where the one-headed arrow shows predictor development follows criterion development.

1.3.5. Predictor Development

Predictor development involves choosing the types of assessments used to measure the selection KSAOs (see arrow 6 in Figure 3). The job analysis may conclude with a blueprint identifying the selection KSAOs, but there is considerable latitude in how to measure those KSAOs in selection.

There are several assessment methods that can be used to measure KSAOs in selection and classification. Each type of assessment method produces scores that may be used for multiple purposes. Likewise, different assessment methods may be used to provide scores that offer inferences about the same KSAO construct. Table 6 lists common examples of assessment methods that differ in their delivery mode. These include:

- Written assessments
- Computerized assessments (e.g., Internet or mobile device assessments)
- Visual and/or aural assessments (e.g., interviews)
- Performance-based assessments (e.g., assessment centers, simulations)

Table 6. Examples of Assessment Methods and Their Associated KSAO Constructs

	Assessment Methods				
KSAO Constructs	**Written**	**Computer/ Web/ Digital**	**Visual and Aural/Oral**	**Interview**	**Performance-Based**
Cognitive					
• Cognitive Ability	X	X			
• Job Knowledge	X	X	X	X	X
Noncognitive					
• Experience	X	X		X	
• Biodata	X	X	X	X	
• Personality	X	X	X	X	X
• Motivation				X	X
• Interests/ Values	X	X	X	X	X
Performance-Based					
• Physical Ability					X

Note: Xs indicate common ways of measuring each construct.
Note: Adapted from Ployhart et al., 2006.

Table 6 shows that different assessment methods are better suited for assessing different KSAO constructs than others. The modality by which KSAO assessments are administered, delivered, and scored can further be understood in several more specific components. According to Lievens and Sackett (2017, p.17), these assessment modes are as follows:

- Stimulus format is the "Modality by which test stimuli (information, questions, prompts) are presented to test-takers"
- Contextualization is "The extent to which a detailed context is provided to test-takers"
- Stimulus presentation consistency is the "Level of standardization adopted in presenting test stimuli to test-takers"
- Response format is the "Modality by which test-takers are required to respond to test stimuli"
- Response evaluation consistency is the "Level of standardization adopted in terms of evaluating test-takers' responses"
- Information source is the "Individual responding to the test stimuli"
- Instructions are "The extent to which directions are made explicit to test-takers about which perspective they should take to respond to the test stimuli"

Each of these components can influence scores, validity, and reactions to the assessment.

It is important to understand that different KSAO assessment methods can produce scores that provide construct validity evidence to the same KSAO. However, assessments differ in terms of their cost, ease of use, administration, user acceptance, and potential score differences across demographic groups (e.g., race, ethnicity, sex). Table 7 provides an overview and summary of these tradeoffs.

Table 7. Strengths, Weaknesses, and Potential Trade-Offs for Different KSAO Assessments

	Goal					
Predictor	**Effectiveness (Accuracy/ Validity)**	**Efficiency/Cost Effective**	**Time to Assess (Slower is Worse)**	**Engaging Applicant Experience**	**Developmental Feedback**	**Diversity/ Fairness**
Effect Size Benchmarks	Low r = .10 Medium r = .20 High r = .30					Low d = .20 Moderate d = .50 Large d = .80
Cognitive Ability	High	High	Fast	Low	Low	Low
Knowledge and Skill	Moderate	High	Fast	Low	Moderate	Moderate
Personality	Low	High	Fast	Low	Moderate	High
Work Interests, Styles, and Values	Low	High	Fast	Low	Moderate-High	Moderate
Biographical Data	Moderate	High	Fast	Low	Low	Low -Moderate
Fit	Low	High	Fast	Low	Low -Moderate	Moderate
Situational Judgment	Moderate-High	High	Fast	Moderate	Moderate-High	Moderate
Structured Interviews	Moderate-High	Moderate-Low	Moderate-Slow	Moderate	Moderate	Moderate-High
Assessment Centers	High	Low	Slow	High	High	High
Work Samples	High	Moderate-Low	Moderate-Slow	High	High	High

Note: Adapted from Ployhart et al., (2018);; based on data from Ployhart and Holtz (2008); Ployhart, Schneider, and Schmitt (2006); Schmidt and Hunter (1998). Note the effect size benchmarks are based on estimates uncorrected for unreliability and range restriction.

1.4 Techniques for Establishing Evidence of Predictive Relationships

The subsections below describe how to interpret and model scores to establish empirical relationships between KSAO predictors and criteria. For purposes of illustration, we consider a prediction situation where the criterion is training performance (TRAINPERF), and the KSAO predictors are overall cognitive ability (COGABIL), conscientiousness (CONSCIENT), adaptability (ADAPT), and emotion regulation (EMOTION). The Appendix provides SAS code showing how to simulate raw data based on a known correlation structure. The code will generate 500 individual scores on each of the five variables.

1.4.1. Scores and Distributions

Scores. Scores represent the information that is provided by assessments and used to make decisions and form inferences about a candidate. All assessments used in selection and classification should produce scores that provide as much information as possible. Scores may be classified in terms of the richness of information they provide:

→ Least Information
- Nominal (categorical information; e.g., male/female, yes/no, hire/reject).
- Ordinal (ordered categorical information but not equal distances between ranks; e.g., rank orders).
- Interval (continuous equidistant information about ranks; e.g., survey response on a five-point least likely to most likely scale).
- Ratio (continuous information with a true zero point; e.g., dollars, biomedical/physical indices).

→ Most Information

Use scales of measurement that are at least interval, whenever possible. First, greater information enables more fine-grained distinctions between candidates. Second, it is possible to use more sophisticated analytical techniques to employ predictive modeling. Third, scores that provide more information tend to be more reliable and manifest more variability (which is necessary for using inferential statistics). See Ployhart et al. (2018).

Distributions. Most KSAOs tend to follow normal (bell-shaped) distributions. Performance distributions are also usually shaped like a bell curve, but there are some instances where truly exceptional performers may produce a power law curve (O'Boyle & Aguinis, 2012). Likewise, some performance results may naturally manifest a skewed distribution because they are low base rate events (e.g., sabotage). These instances are likely to be rare in most USAF roles, and hence assuming normal distributions is most warranted based on the evidence currently available. It is more common that when KSAOs or performance scores differ from normal distributions, the results are usually due to some external influence (e.g., situational factors that restrict the range of scores). The important point to note about distributions of variables is that any deviation from a normal distribution will attenuate effect sizes, such as correlation coefficients. For example, a heavily skewed distribution on performance will attenuate a validity coefficient relative to what would be observed if performance was normally distributed.

1.4.2. Estimates of Relationships

Selection and classification are fundamentally concerned with establishing relationships between KSAO predictor scores and criterion scores. Figure 1 shows this graphically, with arrow 5 representing an empirical relationship between the predictor-criterion scores (arrow 6 is the key inference that is, in part, based on support via arrow 5). The statistical estimate of this relationship is known as an *effect size*. Correlation and regression (i.e., the standardized regression weights or beta-weights) are the two most common ways to estimate effect size. Note the following conventions:

- The scores on the KSAO predictor are denoted by the letter "X"
- The scores on the criterion are denoted by the letter "Y"

Correlation. There are many ways to empirically estimate and test the magnitude of relationships between two variables. Which type of correlation is appropriate is determined by whether the predictor and criterion scores are nominal (categorical) or interval (continuous). There may be situations where the criterion is nominal, such as retention (0 = leave; 1 = retained), and thus a point-biserial correlation is most appropriate. This report focuses primarily on interpreting the Pearson product-moment correlation, as it is the relationship that has seen the most research. Table 8 shows the different types of correlations that can be used for different combinations of nominal and interval data. Note that all of these statistics are intended to estimate a relationship, and hence any of them can be used to support arrow 5 in Figure 1.

Table 8. Correlation Types for Nominal (Categorical) or Interval (Continuous) Relationships

			Criterion	
		Nominal	***Ordinal***	***Interval***
	Nominal	Phi		Point-biserial
Predictor	***Ordinal***		Spearman rank order	
	Interval	Point-biserial		Pearson correlation

The following summarize characteristics of the correlation (we focus on the Pearson correlation, given its widespread use):

- The correlation provides a single number estimate of the (a) strength and (b) direction of a linear relationship
- Range is from
  - -1.00 (perfect inverse relationship; a 1 unit decrease in Y for every 1 unit increase in X; or vice versa)
  - +1.00 (perfect positive relationship; a 1 unit increase in Y for every 1 unit increase in X; or vice versa)
  - 0.00 indicates no relationship
  - The stronger the (absolute) number, the stronger the relationship

Using the sample data and context, the correlation matrix is presented below (see Appendix A for the code to estimate the correlations). Note the top number is the estimate, and number below is the estimate of statistical significance (i.e., *p* value, where $p < .05$ is statistically significant). In this example of 500 candidates, all correlations are statistically significant. The correlations in the box represent criterion-related validity, because they are the correlations between the criterion (training performance) and the KSAO predictors (cognitive ability, conscientiousness, adaptability, and emotion regulation).

	TRAINPERF	COGABIL	CONSCIENT	ADAPT	EMOTION
TRAINPERF	1.00000	0.31880	0.20552	0.23102	0.09956
		<.0001	<.0001	<.0001	0.0260
COGABIL	0.31880	1.00000	0.18994	0.28990	0.10428
	<.0001		<.0001	<.0001	0.0197
CONSCIENT	0.20552	0.18994	1.00000	0.42046	0.13439
	<.0001	<.0001		<.0001	0.0026
ADAPT	0.23102	0.28990	0.42046	1.00000	0.54200
	<.0001	<.0001	<.0001		<.0001
EMOTION	0.09956	0.10428	0.13439	0.54200	1.00000
	0.0260	0.0197	0.0026	<.0001	

Regression. Correlations estimate the relationship between two variables, but selection and classification usually involve a battery of KSAO scores intended to predict a criterion. Indeed, performance is determined by multiple KSAOs, and job analysis rarely identifies a single KSAO predictor of performance. Regression is a highly effective and flexible analytic tool that can be used to estimate the relationship between a set of KSAO predictors and a single criterion. In regression, the KSAO predictors may be continuous or categorical, but the criterion must be continuous. When the criterion is nominal (e.g., retained vs. quit) or ordered categorical, logistic or ordinal regression is appropriate (see Cohen, Cohen, West, & Aiken, 2003, for predicting other types of outcomes).

The basic multiple regression model is shown below:

$$Y = b_0 + b_1(X_1) + b_2(X_2) + \ldots + b_k(X_k) + e$$

Where:

k = number of KSAO predictors
Y = criterion score
b_0 = intercept; or the score on the criterion when the predictors are equal to zero
b_k = regression weight associated with KSAO predictor X_k; how much of a change in Y is associated with a 1 unit change on the KSAO predictor
e = error or residual term

Regression is based on the correlation, but regression can include multiple KSAO predictor variables in the model. Estimating a regression coefficient (denoted by *b*) in multiple regression requires three pieces of information:

1. total relationship of the KSAO predictor with the criterion
2. unique relationship of the KSAO predictor with the criterion
3. correlations among KSAO predictors

Estimating these sources of variance becomes difficult when (a) the KSAO predictors are highly correlated, and/or (b) there are a large number of KSAO predictors. Both situations can contribute to multicollinearity, which makes it difficult to interpret regression coefficients. Therefore, the number of KSAO predictors included in the initial model should be limited based on substantive considerations, practical constraints, and statistical realities (too many predictors that are highly correlated will cause the model to crash). There are no hard rules of thumb, but experience suggests including more than 10-15 KSAO predictors in a model at one time can create statistical and interpretative issues. The main factors that determine how many predictors can be included in one model is primarily based on the correlations among the predictors and sample size. Note that if multicollinearity is an issue, there are many potential solutions, including reducing the number of predictors in the model, or combining the most highly correlated predictors into a composite (see Cohen et al., 2003, for other suggestions).

Regression is helpful in determining:

- **How to combine KSAO predictor scores in the statistically optimal manner**. Regression optimally weights each KSAO predictor according to its relationship with the criterion, balanced against the interrelationships among the other KSAO predictors. The regression model thus provides a weight, or a regression coefficient that estimates how much the KSAO predictor contributes to the overall prediction of the criterion. Assuming equal inputs, the absolute best a human can do in optimally combining predictor information is what standard regression does as a standard practice. Using the sample data, the results of the regression model with four predictors is shown below:

```
                              The REG Procedure
                                Model: MODEL1
                          Dependent Variable: TRAINPERF

                     Number of Observations Read         500
                     Number of Observations Used         500


                              Analysis of Variance

                                       Sum of           Mean
Source                     DF          Squares         Square    F Value    Pr > F

Model                       4         66.12339       16.53085      18.89    <.0001
Error                     495        433.16332        0.87508
Corrected Total           499        499.28671


                Root MSE              0.93546    R-Square     0.1324
                Dependent Mean       -0.02301    Adj R-Sq     0.1254
                Coeff Var         -4065.37628
```

In this example, the overall model R^2 is .13 and statistically significant ($p < .0001$). This means that 13 percent of training performance is explained by the four predictors as a set. The adjusted R^2 is sometimes also helpful in interpreting the overall fit of the model. The adjusted R^2 is the R^2 but "adjusted" downward as more predictors are added to the model. In this sense, the adjusted R^2 rewards parsimony. One can compare the R^2 and adjusted R^2 as a means to see how much the explained variance is being accounted for by the number of predictors in the model.

- **How to use the KSAO predictor information to create a predicted score for the criterion**. The regression model can combine the KSAO predictor information in an optimal manner to estimate a predicted score on the criterion. This is important in selection and classification because one can estimate a prediction equation on a sample of incumbents (concurrent validation) and then use the prediction equation to estimate scores for candidates who have not yet been selected or classified. The equation below shows the generic form of the regression prediction model. Notice Y-hat ($\hat{Y}$) represents the predicted value for the criterion, and the residual term is removed from the model (because the predicted value is based on only the explainable variance).

$$\hat{Y} = b0 + b1(X1) + b2(X2) + \ldots + bk(Xk) \qquad (1)$$

One can use the regression model to provide estimates of each person's predicted score. Using the sample data, the multiple regression model using unstandardized weights is estimated as follows (with SAS output shown):

$$\hat{Y} = -.02 + .27(COGABIL) + .11(CONSCIENT) + .11(ADAPT) - .00(EMOTION)$$

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	-0.02056	0.04187	-0.49	0.6237	0
COGABIL	1	0.27059	0.04459	6.07	<.0001	0.26670
CONSCIENT	1	0.11270	0.04802	2.35	0.0193	0.10940
ADAPT	1	0.11244	0.05824	1.93	0.0541	0.10873
EMOTION	1	-0.00184	0.04899	-0.04	0.9701	-0.00188

One can use the weights as a means to combine the KSAO predictors to estimate a predicted criterion score. For example, assume each KSAO predictor is scored on a 100-point scale (higher numbers are better), and a candidate scores as follows: 72 cognitive ability, 80 conscientiousness, 43 adaptability, and 55 emotion regulation. The regression equation can be used to weight these scores to the optimally predicted training performance score of 32.95:

$$\hat{Y} = -.02 + .27(COGABIL) + .11(CONSCIENT) + .11(ADAPT) - .00(EMOTION)$$
$$\hat{Y} = -.02 + .27(72) + .11(80) + .11(43) - .00(55)$$
$$\hat{Y} = -.02 + 19.44 + 8.80 + 4.73 - .00$$
$$32.95 = -.02 + 19.44 + 8.80 + 4.73 - .00$$

- **Which KSAO predictors are most important in explaining variance in a criterion score**? The regression model can be examined to identify which KSAO predictors contribute most strongly to the prediction of the criterion. The SAS output provides both the unstandardized and standardized regression coefficients. When the scales of measurement differ greatly on the predictors, it is often helpful to examine the standardized estimates. Standardized estimates are calculated by taking the unstandardized estimate times the ratio of the standard deviation of the predictor divided by the standard deviation of the outcome. The unstandardized estimates are interpreted in terms of the original scale units (i.e., how much of a change in Y given a one unit change in X) while the standardized estimates are interpreted in terms of standard deviation units (i.e., how much of a change in Y standard deviation units given a one standard deviation change in X). However, the same statistical significance tests are used for both unstandardized and standardized estimates, and hence the tests are the same regardless of standardization (Cohen et al., 2003).

  To determine which KSAO predictors are most important, one can rely on the effect size (i.e., the regression weight estimate) and the statistical significance of each weight. Using the sample data and results, one sees that cognitive ability is the strongest predictor of training performance (b = .27, p < .05). Conscientiousness (b = .113) and adaptability (b = .112) are nearly identical; b's ≈ .11, but conscientiousness is significant at $p < .05$ while adaptability is not (p = .054). This difference is negligible and likely not practically significant. Emotion regulation has an effect size of b = 0.00 and is not statistically significant (p = .97). Hence, based on these results, cognitive ability is the strongest predictor, followed by conscientiousness and then adaptability (although the difference between these latter two variables is trivial).

Note that in this example the focus has been on the standardized estimates. Whenever the predictors differ on their scales of measurement, it is preferable to examine the standardized estimates because they are more directly comparable. Further, predictors weight themselves by their variance, and so when the measurement scale is arbitrary (as is typical in psychological measurement), it is helpful to focus on standardized estimates to enable more meaningful comparisons.

There are other, more complex ways to estimate KSAO predictor importance. These include use of relative importance methodologies (Johnson, 2000), examination of partial and semipartial correlations, testing subsets of predictors, and so on. Oftentimes these alternative methodologies are consistent with the standardized regression coefficients, but as the number of predictors and/or multicollinearity increases, the estimates from the different approaches can diverge. If there is a desire to estimate relative importance, it may be worthwhile to employ several different approaches to examine the magnitude of potential differences. These approaches are described in detail in Cohen et al. (2003).

- **Which KSAOs are redundant with other KSAOs and hence can be eliminated from the selection and classification model**. Regression models are helpful in reducing the set of potential KSAO predictors to only those that contribute uniquely to the prediction of the criterion. Including too many predictors with only trivial relationships to the criterion is both costly and time consuming. Avoid including KSAO predictors that are redundant with other predictors. Thus, the key is to identify the subset of predictors that are uniquely related to the criterion. One approach to eliminating redundant or irrelevant predictors is to compare regression models. Theory and/or practical considerations can be used to create the alternative models. The baseline model includes the starting set of predictors, and the reduced model includes the smaller set of predictors. Compare the difference in model R^2s; if the difference is not significant than the reduced model should be preferred.

  Continuing the example, a series of "reduced" models are compared to the "full" baseline model 1. This is done simply to illustrate whether different subsets, representing different predictors, explain similar amounts of variance as the full baseline model. However, one could compare the reduced models to each other, such as illustrated in model 5. In the table below, one can see that emotion regulation contributes nothing to the prediction of training performance, so it can easily be eliminated without loss. The difference in R^2s is not significant:

Model	R^2	Compare	ΔR^2	Δ F test
1. COGABIL+CONSCIENT+ADAPT+EMOTION	.1324*	-		
2. COGABIL+CONSCIENT+ADAPT	.1324*	1-2	0.0000	0.00
3. COGABIL+CONSCIENT	.1234*	1-3	0.0100	1.71
4. COGABIL+ ADAPT	.1226*	1-4	0.0100	1.86
5. COGABIL	.1016*	2-5	0.0218	6.17*

* $p < .05$

Thus, we should clearly remove emotion regulation from the model because it adds nothing to the prediction of training performance. We should definitely keep cognitive ability in the model. After including cognitive ability, it does not matter much whether one includes conscientiousness (model 3) or adaptability (model 4) as they are highly similar. However, at least one of them (conscientiousness or adaptability) should be included. If theory and resources were supportive, it would be best to include both conscientiousness and adaptability along with cognitive ability. If there are resource constraints, then one could likely include only cognitive ability and conscientiousness, as conscientiousness is slightly more predictive than adaptability.

To this point the regression model has been focused on predicting continuous criterion scores. However, a form of regression can also be used to model dichotomous outcomes such as re-enlistment or retention. Regression for dichotomous scores is known as *logistic regression*. Logistic regression uses a different estimation method (maximum likelihood), tests of regression weights (Wald's chi-square) and different tests of model fit (AIC, SIC, Wald's, and Likelihood Ratio tests). The parameter estimates give the probability of the outcome occurring given the predictor KSAO. Multinomial and ordinal logistic regression are extensions of the basic logistic regression model used for multiple categories or ordered categories, respectively. See Cohen et al. (2003) for more description of these models. However, the basic application of logistic regression is identical to the regression models summarized in this section and below.

1.4.3. The Necessary Role of Judgment

One should never blindly rely on the results of statistics to inform the design of selection and classification models. Rather, judgment and statistics should work hand-in-hand in designing selection and classification systems. First, judgment (informed by theory, experience, and practice), should be used to guide the choice of statistics (e.g., which measures to use, which predictors to include, nature of criteria, etc.). Second, statistics should then be employed to estimate, test, and combine the scores. Finally, judgment should be used to interpret and make sense of the results. Judgment and statistics thus form a necessary balance of checks and balances.

One of the most common tensions between judgment and statistics is interpreting whether an effect is "meaningful." There are multiple ways to make such an inference:

- The effect size, such as the correlation or regression coefficient (standardized or unstandardized)
- The statistical significance (p value) of the effect size
- The standard error and confidence interval of the effect size
- Practical judgments about what is considered a small, medium, or large effect

Remember that effect size, statistical significance, sample size, and statistical power are all interrelated (Murphy & Myors, 1998). Statistical significance is heavily influenced by sample size, as larger samples produce smaller standard errors and greater statistical power. The conventional level of statistical significance is that an effect "is significant" if $p < .05$. However, one should not interpret statistical significance unless there is sufficient statistical power to test the effect. When making high-stakes decisions, as in selection and classification, you want

statistical power to be as high as possible. A historical rule of thumb is statistical power must be at least .70, but this is too low for selection and classification. Unless practical constraints prohibit it, statistical power should be .90 or higher. If practical constraints prohibit a sufficient sample to get to power of .90, then consider raising the *p* value (e.g., $p < .10$). Power should be estimated on the effect size of interest (e.g., a regression weight, a model R^2, an incremental R^2, etc.). Murphy and Myors (1998) provide a very simple but useful framework for estimating statistical power across a variety of situations and effect sizes.

Note that just because an effect size is statistically significant does not mean it is practically significant. With very large sample sizes (e.g., over 10,000 observations, over 100,000 observations), even trivial effect sizes ($r = .0001$) are statistically significant. Likewise, in smaller sample sizes ($n = 20$), even large effect sizes ($r = .40$) may not be statistically significant). Importantly, effect size estimates are not affected by sample size (although their confidence interval will be affected).

The approach advocated here is to use multiple pieces of information to make the best use of judgment and statistics (see Murphy & Jacobs, 2012). The following approach is helpful to guide decision making:

1. Consider the sample size and statistical power to test the effect size.
    a. If sufficient, then interpret the statistical significance to determine if $p < .05$
        i. If $p < .05$ or reasonably close, examine the effect size estimate
        ii. If $p > .05$, consider the effect as ignorable
    b. If insufficient, consider collecting more data (if possible) or raising the p value (if collecting more data is not possible)
2. Balance interpretation of statistical significance with practical significance. The determination of practical significance is based on a priori values and judgment. Consider a statement such as: "To be included in a predictor battery, a KSAO predictor must have a regression weight that is statistically significant and explains at least .05% of the variance in the criterion."
    a. It is often difficult to determine what is practically significant. If the criterion scores have a metric that is inherently meaningful, such as number of accidents, then one could determine how many accidents would have to be reduced by using the predictor for selection, as a means to set practical significance. For criteria such as training grades or attrition, one can use the same logic. Organizational leaders can help identify what they consider to be practically meaningful (perhaps based on operational metrics, cost, etc.), and then interpret effect sizes relative to those standards of practical significance.
    b. Remember, the regression weight represents how much of a change in the criterion is associated with a one unit change in the predictor. If one believed that reducing 8 accidents was minimally necessary for incurring the cost of a predictor, then one would require the regression weight to be statistically significant ($p < .05$) and be at least $b = -.08$ or greater.
3. The key is to never rely solely on human judgment or statistics. A balanced approach ensures the best overall quality decision making.

1.5 Artifacts and Conditions Affecting Predictive Relationships

There are several factors that influence the estimation of effect sizes (e.g., correlation, regression coefficients). These are often called "artifacts" and need to be considered (or addressed using various artifact corrections), as appropriate. The following artifacts will attenuate (reduce) effect sizes, and thus reduce statistical power and statistical significance as well.

1.5.1. Scale Coarseness

In the earlier section on scores, it was noted that more information is provided by interval than nominal scales. More information is provided when interval or ratio scores are used because they enable more fine-grained distinctions. This may be understood more generally as *scale coarseness*, which refers to how much information is provided by the measurement of a KSAO or performance construct. Coarse scales have fewer data points, while fine scales have more data points. For example, a 7-point scale is finer than a 3-point scale. In general, more coarse scales prohibit more nuanced distinctions among scores, and will, all else being equal, produce smaller effect sizes. One should never dichotomize a continuous score, as doing so reduces effect size and hence statistical power.

However, this does not mean that more fine scales are inherently better. Using a self-report measure of job satisfaction with a 1,000-point scale will not produce more meaningful variance and may actually increase error variance. The appropriate number of scale options (e.g., 3 point, 5 point, 7 point, etc.), should be based on theory and the number of options respondents actually endorse. Use the finest scale that respondents will actually use and understand.

This is a common issue in performance ratings. Raters are often reluctant to give unfavorable scores (i.e., leniency bias), so one may believe that providing more scoring options will fix the problem (e.g., using a 9 point scale instead of a 5 point scale). The reality is that adding more response options will only matter if raters actually use all response options. If raters only make low, moderate, and high distinctions between employees, then a 3-point scale may actually be sufficient.

The bottom line is that one should use the most fine-grained scale as realistic and practically feasible. Use of BARS will further help ensure adequate ratings, and rater training will ensure raters understand how to use the scale. Never reduce a continuous score except in extremely rare circumstances.

1.5.2. Nonnormality

Correlation and regression assume normal distributions of the KSAO predictors and criteria. Although they are fairly robust to violations of normality, any deviation from normality of the variables will reduce effect sizes and statistical power. Therefore, anything that distorts a normal distribution—rater leniency or severity, faking on noncognitive assessments, etc., will attenuate effect sizes. If distributions are severely nonnormal (rules of thumb suggest skew greater than 3 and kurtosis greater than 7 in absolute numbers), then one must consider whether the distribution should be transformed or a different analytic approach should be used. For example, one could transform a nonnormal distribution into one that is more linear. If the data are heavily skewed,

one could use a square root transformation. In some situations (studying "star" performers who create a disproportionate impact), a power law curve may be the more appropriate distribution (Aguinis & O'Boyle, 2014), and in such situations a different statistical model (such as a generalized linear model) could be employed. A review of transformations in the organizational literature is provided by Becker, Robertson, and Vandenberg (2018). Of course, one must also consider the distribution of the residuals when conducting regression analyses, as another means to assess the fit of the model (see Cohen et al., 2003).

1.5.3. Nonlinearity

Correlation and regression assume linear relationships among KSAO predictors and criteria. There are many instances where a relationship may be nonlinear (e.g., the relationship between group task conflict and group performance). However, the estimates of correlations or regression coefficients will be attenuated as the relationship is nonlinear. Hence, slight deviations from linear relationships will not be very noticeable, while strong nonlinear relationships will greatly attenuate effect sizes. An example of the latter is when there is a perfect u-shaped (or inverted u-shaped) relationship—a correlation will be zero in such a situation. Hence, always examine whether the relationship is linear by examining diagnostics (e.g., residual diagnostics such as studentized or standardized residuals), graphs (e.g., scatter plots), and testing for nonlinear via squared terms in regression (see Cohen et al., 2003).

1.5.4. Unreliability

Reliability refers to the consistency of scores, and ranges from zero (perfectly unreliable) to 1.00 (perfectly reliable). As unreliability increases, effect sizes decrease. There is no perfectly reliable score, and unreliability may occur in KSAO predictor scores, criterion scores, and both. One should always estimate reliability. Most frequently this occurs with internal consistency reliability, but other forms may be appropriate (see Schmidt & Hunter, 1996). If one wanted to estimate the correlation correcting for both KSAO predictor and criterion unreliability, one would employ the formula below:

(2)

$$\rho = \frac{\boldsymbol{r}_{xy}}{\sqrt{\boldsymbol{r}_{xx}\boldsymbol{r}_{yy}}}$$

in this equation,

ρ = corrected correlation

r_{xy} = observed correlation between the KSAO predictor (X) and the criterion (Y)

r_{xx} = reliability of the KSAO predictor score

r_{yy} = reliability of the criterion score

However, in selection and classification contexts, it is appropriate to only correct for unreliability in the criterion scores but not the KSAO predictors. The reason is because hiring decisions are necessarily based on fallible KSAO predictor scores. Correcting for unreliability in the criterion thus leads to a simpler formula for correcting for attenuation:

$$\rho = \frac{\boldsymbol{r}_{xy}}{\sqrt{\boldsymbol{r}_{yy}}} \tag{3}$$

Different types of scores tend to manifest different levels of reliability. For example, scores obtained from objective criteria (e.g., absence, accidents, retention) may be measured with almost perfect reliability, whereas scores obtained from subjective criteria (e.g., supervisor or peer ratings) may have lower reliability. Also, different types of reliability tend to provide different estimates. For example, internal consistency reliability may be high (around .70 or larger) while interrater reliability (correlation of ratings made by two raters) is often lower (approximately .50). Finally, if correcting for unreliability for multiple predictors in a regression equation, it is best to correct each predictor for unreliability and then apply the regression model to the corrected coefficients (this is similar to conducting a structural equation model).

1.5.5. Range Restriction

Correlations and regression coefficients are based on variance and covariance. Anything that restricts the variances of KSAO predictor and/or criterion scores, will reduce the effect size. Range restriction is like unreliability; it is pervasive in selection and recruitment models. For example, if one screens on a set of assessments (e.g., AFOQT or ASVAB), assigns candidates to occupations, and then estimates the criterion-related validity on those assigned in the positions, the resulting correlation will be attenuated because only those that passed threshold on the assessment are in the job. This is known as *direct range restriction* and occurs when selection is based on the assessment scores being validated. *Indirect range restriction* occurs when the scores being validated are restricted due to their relationship with a different set of scores that are used for selection and classification decisions (e.g., scores on the TAPAS are not used for making selection decisions, but are nonetheless restricted because they are correlated with ASVAB scores that are used for selection decisions).

Correcting for range restriction requires obtaining estimates of the score variance from an unrestricted sample. This may be best served by obtaining scores from candidates early in a selection process, or collecting data for normative purposes across similar individuals. Practical constraints frequently do not allow much choice about where or when one obtains the unrestricted sample variance. A common way to obtain unrestricted estimates is to examine test norms (if using published assessments). The most important consideration is that the unrestricted estimates are based on samples that are comparable to the one that is restricted, in terms of demographics, experience, etc. If there is a multiple hurdle process (e.g., selection, then classification), it would be desirable to obtain variance estimates at each hurdle.

In the above example, it would require getting variance estimates from the full set of candidates before classification decisions were made. Note range restriction can also occur on the criterion. Table 9 summarizes Sackett and Yang (2000), who provide a typology of range restriction effects and formulae for correcting them. The corrections in Case 1 involve direct range restriction on the predictor (X), criterion (Y), or some combination of the two. For example, direct range restriction on the predictor might occur if candidates for a position had been hired based on an assessment, and then at some later point in time a correlation is estimated between the assessment scores and some criterion. There is direct range restriction on the assessment because only those who scored above a cut score remain in the sample. Case 3 corrections involve what is known as indirect range restriction, where the restricted variance is caused by a third variable. For example, suppose one wants to examine the correlation between cognitive ability (X) and job performance (Y) in an occupation, but access to the occupation was restricted based on grade point average (GPA; Z). GPA is correlated with cognitive ability, and so selecting on GPA indirectly restricts the range on cognitive ability. Case 3 provides estimates that deal with multiple predictors that are used in simultaneous or sequential selection (discussed shortly). Case 4 deals with the difficult situation of trying to estimate unrestricted variance from just a sample. Case 4 is rarely implemented in practice except in an exploratory sense. It is very important to use the appropriate formula when making corrections for range restriction (see Van Iddekinge & Ployhart, 2008).

Table 9. A Typology of Range Restriction Models (Sackett & Yang, 2000)

1. Selection on either x or y; no third variable involved
   a. Unrestricted variance known for selection variable (Thorndike's Case 2)
   b. Unrestricted variance known for other variable only (Thorndike's Case 1)
   c. Unrestricted variance known for neither variable
2. Selection on z
   a. z measured; unrestricted z variance known (Thorndike's Case 3)
   b. z measured; unrestricted z variance unknown
   c. z unmeasured; unrestricted variance known for x and y
   d. z unmeasured
3. Simultaneous or sequential selection
   a. Simultaneous selection on multiple variables; all selection variables measured, and unrestricted variance known for all selection variables (Aitken–Lawley's multivariate case)
   b. Sequential selection on multiple variables; all selection variables measured, and unrestricted variance known for all selection variables (e.g., selection on x, then on y, and then on z)
   c. All selection variables measured, and unrestricted variance not known for one or more selection variables
4. No information about how (or whether) restriction occurred

Reprinted with permission from publisher

1.5.6. Applications and Corrections

The artifacts noted above work in combination to attenuate effect sizes. It is best to try to reduce their influence as much as possible through the use of sound validation designs and assessments that provide highly reliable scores. Unfortunately, it is impossible to eliminate these artifacts in operational contexts. The following factors (already noted) are nearly always present to some degree:

- Individual characteristics
  - Candidate response distortion and faking
  - Candidate motivation
  - Rater biases (e.g., leniency, halo, etc.).

- Measurement and assessment method characteristics (see Table 6 and Lievens & Sackett, 2017)
  - Methods (paper, web, personal device, etc.)
  - Response format
- Contextual characteristics
  - Demand characteristics
  - Instructions
  - Common source bias/method bias (typically inflate effect sizes)

Because it is impossible to eliminate artifacts, researchers will often try to statistically correct for them. This enables the researcher to estimate the observed (attenuated) effect size and the effect size disattenuated from artifacts. The two most common corrections are for unreliability and range restriction. The process involves first correcting for unreliability in the criterion (using the formula shown above), and then correcting for range restriction (using one of the approaches shown in Table 9). See Hunter, Schmidt, and Le (2006) for details, and Van Iddekinge and Ployhart (2008) for a broader discussion of these issues. Note that when using corrections, it is important to always report both uncorrected and uncorrected values, clearly label these values, and describe which corrections were performed and how they were performed.

1.6 Combining Predictor Information for Selection and Classification

It is usually the case that multiple KSAOs will be required to fully understand and predict important performance criteria. Further, using multiple KSAO predictors is usually necessary to simultaneously enhance predicted performance and diversity. However, not all KSAO predictors will contribute uniquely to the prediction of performance. Including unnecessary or redundant predictor KSAOs wastes time, money, and other resources. When done correctly, using multiple KSAO predictors helps balance the strengths of each approach with the weaknesses of the other approaches in a manner that is efficient and effective.

The topic of choosing the best KSAO predictors was introduced in the regression section above. Here the topic is considered further and more broadly from the perspective of designing an effective selection and classification system. When combining predictors into the design of a system, several factors need to be considered:

- Number of candidates being tested
- Cost of each assessment
- Time to administer and score each assessment
- Diversity in candidate pool
- Reactions to assessments
- Acceptance of assessments and results from each major stakeholder group (e.g., leadership, supervisors, candidates, testing specialists)
- Importance of the job
- Correlations among the KSAO predictors
- Validities of KSAO predictors

In turn, these factors influence:

- The choice between speed versus cost versus accuracy
- Use of cut scores and minimal KSAO requirements
- Selection ratio and adverse impact
- Types of assessments used
- The nature of the process (e.g., multiple stages, multiple hurdles, multiple cutoffs, etc.)

These topics are discussed next.

1.6.1. Combining KSAO Predictor Scores

There are two broad approaches to combining predictor scores: compensatory and noncompensatory.

Compensatory. Compensatory models allow the strengths in one type of KSAO to offset potential weaknesses in another KSAO. The regression model is an example of a compensatory way of combining information, as each KSAO predictor is weighted by its contribution to the prediction of criterion scores. Continuing the regression example from above, the prediction equation shown below is compensatory in the sense that a candidate could make up for lower cognitive ability scores by having higher scores on conscientiousness and adaptability. This is illustrated by two illustrative candidates who have the same predicted score. Yet, candidate 1 has performed poorly on the cognitive assessment relative to conscientiousness and adaptability. Candidate 2 has performed well on the cognitive assessment, but poorly on conscientiousness and adaptability. Hence, they have the same predicted score because their relative strengths have helped offset their relative weaknesses.

$$\hat{Y} = -.02 + .27(COGABIL) + .11(CONSCIENT) + .11(ADAPT)$$

Candidate 1: $23 = -.02 + .27(21) + .11(80) + .11(80)$
Candidate 2: $23 = -.02 + .27(70) + .11(20) + .11(20)$

Use of compensatory models is the statistically optimal way to combine predictor information (assuming generalizability and cross-validity). *Using the statistical weights to create predicted criterion scores, rank-ordering candidates from highest to lowest scores, and then selecting in a top-down manner is the most valid way to make selection decisions*. No other linear modeling approach to combining predictor information will outperform the use of a compensatory, regression-weighted prediction model (given that the assumptions for the regression model are met, (e.g., linear relationships) (e.g., Dawes, Faust, & Meehl, 1989; Kuncel, Klieger, Connelly, & Ones, 2013).

Noncompensatory (Cut Scores). Noncompensatory systems are ones where a minimum threshold of competence on one or more KSAOs is required. The system uses a cutoff, where anyone scoring below threshold is eliminated from the selection process. There are many times where cutoffs are required. Examples include physical abilities (e.g., minimum vision requirements for pilots), mental abilities (e.g., high level of cognitive ability needed for advanced training), and credentialing/educational standards (e.g., being certified as an accountant or

passing the Bar as a lawyer). In these situations, it is important to set cut scores that truly reflect the minimum KSAO requirements for being able to perform the job. Cut scores are also frequently used when there are large numbers of candidates because they can make hiring decisions more efficient and timely. Cut scores may also be used when the selection process takes a long time to complete. In such cases it may be better to immediately inform those who score low and will never be eligible, so that they can pursue other opportunities.

Setting cut scores means identifying a threshold on KSAO scores where falling below the value means one cannot do the job. Setting cut scores is inherently operationalizing a value judgment of what is considered acceptable and unacceptable. Setting cut scores means one needs to identify the minimum KSAO level for a person who can competently perform the job—that is, a minimally competent person (MCP). As noted above, one should generally avoid setting cut scores unless truly necessary, because setting cut scores dichotomizes an otherwise continuous distribution (and thus reduces effect size and statistical power). However, when it is required for safety or practical reasons, the cut score must be set based on solid theoretical and empirical evidence because the cut score has important consequences (APA Standards, 2014).

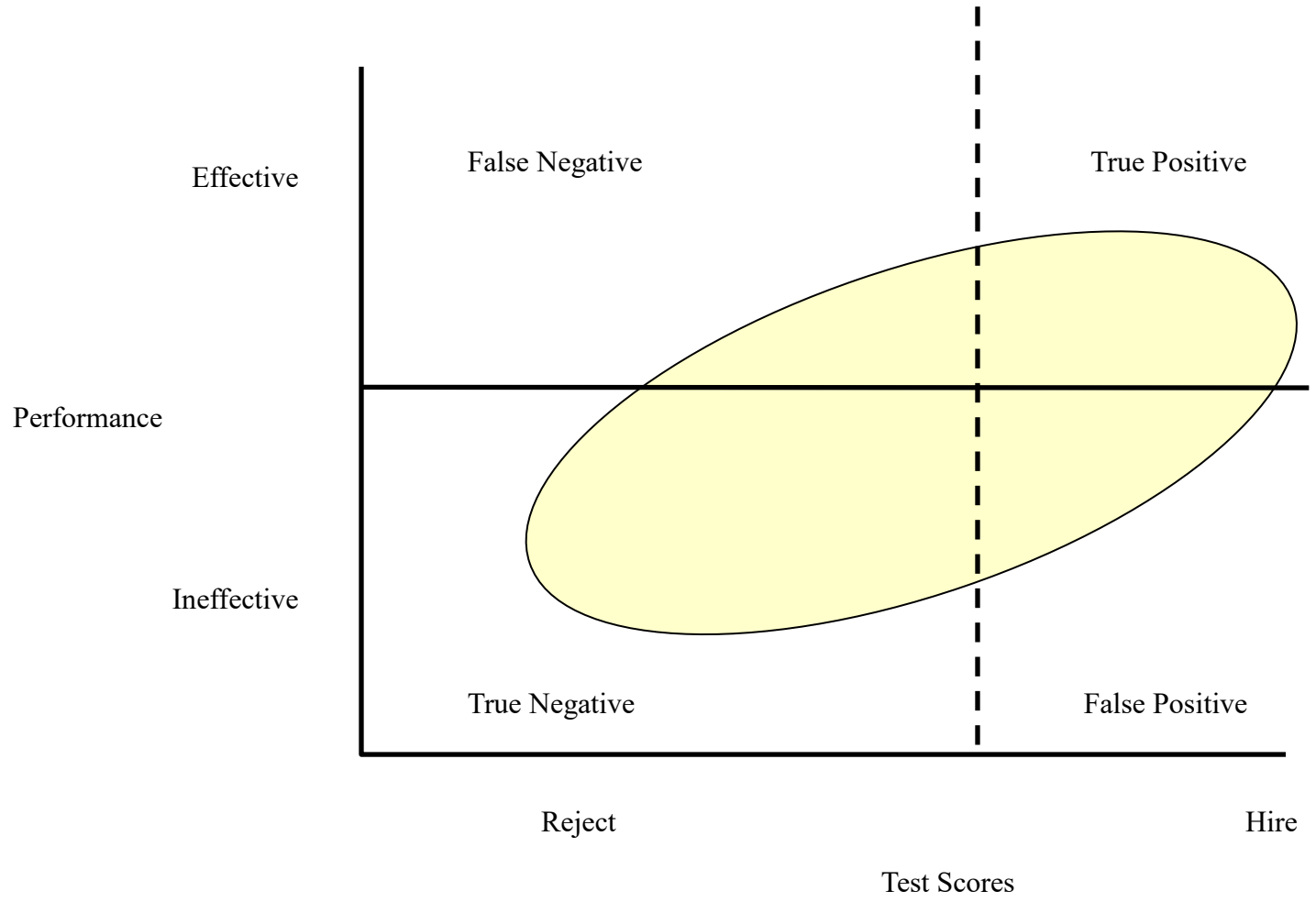
Figure 4 illustrates the consequences of cut scores. The dashed line represents the cut score and can be moved to the right (setting a higher cut score) or left (setting a lower cut score). For purposes of illustration, the performance distribution is dichotomized into effective versus ineffective. The ellipse represents the scatterplot of scores or the relationship between test scores and performance (in this example, the relationship is positive). The two lines dissect the scatterplot into four quadrants.

- *True positives* are those who score above the cut score and can effectively perform the job.
- *False positives* are those who score above the cut score but cannot do the job. They incorrectly passed the cutoff (there is no perfect prediction), but it turns out they were unable to effectively perform the job.
- *True negative*s are those who scored below the cutoff and cannot do the job.
- *False negatives* are those who scored below the cutoff, but could have effectively performed the job (if allowed the opportunity to do so).

Figure 4 is obviously theoretical (it's usually impossible to know who is a false negative), but the figure is instructive in illustrating the consequences of cut scores.

- Raising the cut score reduces the number of people selected, but increases the odds that each person selected will be able to effectively perform the job (true positives).
- Lowering the cut score increases the number of people selected, but increases the odds that some selected will not be able to effectively perform the job (false positives).

The level of the cut score should correspond to the consequences of error in the job. For extremely critical jobs with high consequences (e.g., pilots of nuclear capable aircraft), high cut scores are needed and higher false negative rates should be tolerated.



Note. The dashed line represents the cut score.

Figure 4. The Effects of Cut Scores on Classification Decisions and Statistical Power

Setting cut scores thus requires validity evidence just as another part of selection and classification model design (SIOP Principles, 2018; APA Standards, 2014). There are several approaches to setting cut scores (Guion, 2011; Ployhart et al., 2006).

- **Performance-based.** Examine criterion performance distributions for different groups of incumbents who also have scores on the KSAO predictor (e.g., those judged to be ineffective, minimally effective, or superior). Using the criterion score distribution, identify the point where performance is considered minimally effective. Now, using a scatter plot, identify the score in the corresponding KSAO predictor distribution that matches the criterion score—this point becomes the cut score.
- **Expert judgment.** SMEs can be used to set the cut score. There are a variety of approaches, but one popularized by Angoff is most common (see Jaeger, 1989). First, SMEs examine each KSAO assessment item and identify the percent of MCPs who could answer the question correctly. Second, average the percentages for each item, being sure to evaluate whether there is sufficient consensus. Third, the average percentage for each item is averaged across the full range of items. This overall average percentage represents the minimum passing score (e.g., 75% of items correct), and thus establishes the cut score.
- **Regression-based.** When there are multiple predictors, it is sometimes easier to set cut scores using regression. Regress the criterion score onto the predictors to estimate the regression weights. The predicted value associated with minimally acceptable performance becomes the cut score. Note this approach uses all of the information from the predictors in a compensatory manner, but then creates a cut score based on the overall composite score.
- **Item Response Theory-based.** Item Response Theory (IRT) models are modern approaches to item and test construction, and are generally preferable to classical test theory approaches. IRT item and test characteristics can be used to provide a greater level of precision for passing the item and test than the expert judgment approach noted above.

Choosing Between Compensatory and Noncompensatory Models. There are advantages and disadvantages to compensatory and noncompensatory models. Generally, a compensatory approach is to be preferred because it uses the most information in the statistically optimal manner. That is, the predictor information is combined via regression, and then the cut score is based on the overall composite score. This compensatory approach is particularly effective in situations such as the Air Forces, where individuals must have information regarding which of a multitude of jobs they are qualified for. The practical factors that lead to the choice of one approach over the other should be clearly stated (SIOP Principles, 2018).

The advantages of a compensatory system:

- Maximizes prediction (top-down selection)
- Optimal way to combine KSAO predictor information
- Flexible to different types of predictor KSAOs

The disadvantages of a compensatory system:

- Requires large sample sizes
- Must assess all candidates on all KSAO predictors

The advantages of a noncompensatory system:

- Increases odds those hired are minimally qualified
- Necessary for credentialing, licensure, etc.
- Can save time and money (e.g., can inform large numbers of candidates who will not be hired)

The disadvantages of a noncompensatory system:

- Requires large sample sizes
- Establishing validity of cut score can be difficult and time consuming
- Dichotomizing continuous scores reduces effect size and statistical power

Statistical Versus Unit Weights. When combining multiple KSAO predictor scores, there are multiple ways that the scores can be combined (Bobko, Roth, & Buster, 2007; Johnson & Oswald, 2017; Oswald, Putka, & Ock, 2015; Sackett, Dahlke, Shewach, & Kuncel, 2017). Theoretical, practical, and empirical considerations should be given to this choice, and the rationale clearly described. Among the major choices:

- **Unit weights.** Instead of using regression weights, each KSAO predictor score is simply summed along with other scores. For example:

  $$\hat{Y} = 1(\text{COGABIL}) + 1(\text{CONSCIENT}) + 1(\text{ADAPT}).$$

  This unit weight approach works surprisingly well, can frequently generalize across contexts better than other approaches, is easy to apply and explain, and will correlate highly with empirical approaches so long as the signs of the regression weights are the same as the unit weights (Bobko et al., 2007; Dawes, Faust, & Meehl, 1989; Kuncel, Klieger, Connelly, & Ones, 2013). However, keep in mind that regression-based estimates will be more precise than unit weights, so the choice between unit weighting and regression weighting is heavily dependent on the practical use of the scores..
- **Rational weights.** SMEs use judgment to estimate the weights they believe are appropriate for each KSAO predictor. These judgments are frequently wrong, as even experts can be quite poor at combining multiple types of information. There are many opportunities to unintentionally make mistakes using this approach (Oswald et al., 2015).
- **Correlations.** Bivariate correlations are used to weight each KSAO predictor. For example:

  $$\hat{Y} = .32(\text{COGABIL}) + .21(\text{CONSCIENT}) + .23(\text{ADAPT}).$$

  This approach is problematic and will result in less optimal prediction than regression because it fails to take into consideration the intercorrelations among the predictors.
- **Regression weights.** They are the optimal way to combine KSAO predictor information, in no small part because they account for KSAO predictor intercorrelations (Dawes et al., 1989). For example:

  $$\hat{Y} = .02 + .27(\text{COGABIL}) + .11(\text{CONSCIENT}) + .11(\text{ADAPT}).$$

  However, weights may be sample-specific, so it is important to estimate the cross-validity of the weights (discussed shortly). The other challenge to regression weights is multicollinearity. Mild forms of multicollinearity are usually not a serious problem. However, when there is high multicollinearity, it becomes difficult to provide stable regression weights and can also increase the standard error of the weights. Sometimes the

signs of the predictors (positive or negative) will flip when changing predictors in the model. In these situations one must be cautious about applying the regression weights, and additional steps should be undertaken to examine the robustness of the weights (e.g., create composites of highly-correlated predictors, remove redundant predictors, see Cohen et al., 2003).

- **Pareto weights.** In contrast to regression weights, which maximize the prediction of a single criterion, a pareto weights approach is used when seeking to maximize the prediction of two criteria. They have been employed as a means to weight a KSAO predictor composite to balance the prediction of performance and diversity outcomes (e.g., De Corte, Lievens, & Sackett, 2007). These approaches are promising, but more needs to be learned about their operational use and generalizability (Song, Wee, & Newman, 2017), especially given their complexity and the mixed evidence fully supporting their use.

1.6.2. Types of Selection and Classification Systems

Selection and classification systems can be administered in a number of different ways. There are five broad approaches.

- **Single stage.** Single stage systems administer all of the KSAO assessments used for making selection decisions in a single administration. These systems are fairly rare in practice, as it is common to put basic screening tools (e.g., information about eligibility, such as meeting basic age requirements or basic physical readiness) first in the process to ensure only eligible candidates are assessed.
  - Compensatory. Single stage compensatory systems administer one or more KSAO assessments and use the full range of scores on each assessment to make an overall selection decision.
  - Noncompensatory. Single stage noncompensatory systems administer one or more KSAO assessments, but there are cutoffs placed on one or more assessments that must be met. Those who pass the cut scores will have their KSAO scores combined while those who fall below the cut score are eliminated from the process.
- **Multistage.** Multistage systems use one or more stages where different KSAO assessments are administered. For example, a basic readiness screen may be first provided to ensure potential candidates meet basic qualifications (e.g., age for enlistment eligibility). Then, low-cost and easy to administer assessments may be used (e.g., ASVAB or AFOQT assessments). Next, more intensive and costly assessments (e.g., simulations, interview) may be used to provide the most depth into the candidate's qualifications.
  - **Compensatory.** Multistage compensatory systems have several distinct stages where candidates complete different KSAO assessments, but all candidates will go through all stages. The scores at each stage are then used to provide an overall selection decision.
  - **Noncompensatory (Sequential).** Multistage sequential noncompensatory systems have several distinct stages where candidates complete different KSAO assessments. Cut scores may be used at each stage, but all candidates will go through all stages. The scores at each stage are then used to provide an overall selection decision.

However, candidates who score below the cut score on any of the relevant assessments will be eliminated from the process. The benefits of this approach are that it provides complete information on all candidates and gives all candidates an opportunity to perform all stages. The disadvantage of this approach is that it requires assessing all candidates, including those who cannot perform the job.

- **Noncompensatory (Hurdle).** Multistage hurdle noncompensatory systems have several distinct stages where candidates complete different KSAO assessments. Cut scores may be used at each stage, and only candidates who score above the cut score will proceed to the next stage. For the candidates who pass all stages, the scores at each stage are used to provide an overall selection decision. The benefits of this approach are that it saves time and money because only those candidates who pass the cut score move to the next stage. The disadvantage of this approach is that it requires justifying the cut scores at each stage and those candidates may not feel they had a full chance to demonstrate their abilities.

1.6.3. Additional Practical Considerations

Among the five broad types of selection and classification systems, there are an enormous number of choices and judgment calls. For example, which predictors should go first in the system, how should scores be combined and weighted, what are the potential consequences on validity and diversity, what are the potential consequences in terms of resources and cost, etc.? Unfortunately, there are no simple rules of thumb that can answer these questions. However, the most critical issues are the selection ratio, incremental validity of later stages relative to earlier stages, and intercorrelations among KSAO predictors across stages (see Sackett et al., 2017). For the Air Force, most questions relating to incremental validity involve examining how much the variance explained increases for a new assessment (e.g., TAPAS) relative to an existing assessment (i.e., ASVAB). It is important also to consider the impact of artifacts such as unreliability and range restriction (Roth, Le, Oh, Van Iddekinge, & Robbins, 2017). Note that finding incremental validity over a long-established predictor like the ASVAB is an important accomplishment. It is very challenging to find incremental validity over most established measures of ability. Incremental validity sets a high, but valuable bar as it ensures any new predictor truly adds value.

It is best to first review the literature on different systems (especially if the system has never been used operationally), and then model different approaches empirically as much as possible. The different approaches may then be compared in terms of validity, diversity, cost, efficiency, etc. These systems and their consequences should be fully explained, and a justification provided for the chosen system (for more details see Aiken & Hanges, 2017; Johnson & Oswald, 2017; Kehoe & Sackett, 2017; Sackett & Ellingson, 1997; Sackett et al., 2017; Tippins, Solberg, & Singla, 2017).

1.7 Subgroup Differences and Adverse Impact

1.7.1. Overview of Diversity and Validity

Candidates into the USAF are diverse in terms of demographic subgroups including sex, race, and ethnicity. It is important to ensure that selection and classification models are fair and

unbiased against members from different demographic subgroups. First, ensuring equal access is simply the right thing to do. Second, failing to ensure equal access violates U.S. federal guidelines regarding merit and equal opportunity.

Fairness in selection and placement is a social perception and differs according to stakeholder values. For example, regardless of any technical merits, some stakeholders may be opposed to any type of assessment while others value such approaches. Bias is a technical and scientific term that, in selection and classification models, refers to situations when there are differences in selection rates across subgroups even when the KSAO scores (i.e., the estimated "true" scores) are equivalent across subgroups. Differences in selection rates that reflect real differences in KSAOs are not bias. Differences in selection rates that occur for reasons other than latent KSAOs are potentially bias.

Subgroup differences refer to whether there are mean and/or variance differences across demographic categories (e.g., race, sex) in KSAO scores and criterion scores. The relevant demographic groups are defined by the U.S. federal government. If there were no subgroup differences in KSAO or criterion scores, then there would be no need to consider the question of diversity and validity separately. However, there are many demographic subgroup differences on different KSAOs, KSAO assessment methods, and criteria, and some of these differences are large enough to negatively impact diversity. Further, some of the most valid KSAO scores (e.g., cognitive ability) produce the largest subgroup differences (i.e., race).

Thus, it is critical to develop selection and classification models that maximize validity and diversity, in a manner that is perceived as fair and is sustainable.

1.7.2. Nature and Consequences of Subgroup Differences

Subgroup differences are defined in terms of mean score differences on KSAOs, KSAO assessment methods, and/or criteria. To ensure appropriate comparisons, mean differences should be expressed in standardized (standard deviation) units. The *d* statistic is used to estimate such differences and is calculated as follows:

$$d = (\text{Subgroup1 mean} - \text{Subgroup2 mean}) / SD_{pooled}$$

$$\text{where, the } SD_{pooled} = \sqrt{\frac{(n_1-1)SD_1{}^2+(n_2-1)SD_2{}^2}{n_1+n_2}} \qquad (4)$$

With this formula, any mean difference between groups is expressed in terms of standard deviation units. Therefore, $d = .30$ says one group scores .30 standard deviation units higher than another group; $d = -.80$ says one group scores .80 standard deviation units lower than another group. It is difficult to provide guidelines for *d* values, but convention suggests small mean differences are $d = .30$ or lower, medium $d = .50$, and large $d = .80$ and higher.

Just because there is a large mean difference does not suggest there will be differences in hiring rates. The selection ratio, defined as the number who are selected over the number who apply, determines the consequences of mean differences on selection rates (Sackett & Wilk, 1994). The hiring rate is the number of people who accept an offer divided by the number of people who

apply. Sometimes the hiring rate is also called a conversion rate, which is the number of people who accept an offer divided by the number of people extended an offer by the organization.

- When there are low selection ratios (few people are hired), small mean subgroup differences in assessment scores can create differences in hiring rates.
- When there are high selection ratios (most people are hired), even large mean differences in assessment scores may not create differences in hiring rates.

Mean differences should always be reported for each relevant group, and the selection ratio for each group and overall. Mean differences should be reported for each KSAO predictor, the criteria, and any composite selection and classification models based on combinations of KSAO predictors.

Based on decades of scientific research, racial subgroup differences tend to increase as the performance task or KSAO predictor is more cognitively loaded (Jensen, 1998). Furthermore, the more the KSAO assessment method requires greater cognition, the more likely there will be racial subgroup differences (Ployhart & Holtz, 2008). For example, assessment center exercises that involve tasks requiring the manipulation of factual information (e.g., in-basket or business information) will have a higher cognitive loading than tasks involving interpersonal activities (e.g., group influence).Note that subgroup differences for criteria tend to be approximately one-third of a standard deviation. Subgroup differences for KSAOs and KSAO assessment methods can be small to large. Generally, the largest subgroup differences are found on cognitive ability assessments, between Asians and Blacks and Hispanics (*d'*s approximately one or greater). The largest subgroup differences are found on physical ability assessments of strength, where men score more than one SD higher than women. Subgroup differences on personality KSAOs are generally small. Table 7 (above) provides a summary of subgroup differences across different types of KSAOs and assessments (see Hough, Oswald, & Ployhart, 2001; Ployhart & Holtz, 2008; Sackett, Schmitt, Ellingson, & Kabin, 2001; for more details).

1.7.3. Adverse Impact

Adverse impact may occur when a selection and classification model is applied consistently, yet produces subgroup differences in hiring rates. Note that the presence of adverse impact is not proof of bias or discrimination; it is only suggestive that bias or discrimination may exist. There are many ways to estimate adverse impact (Bobko & Roth, 2010; Murphy & Jacobs, 2012). The classic approach is to use the 4/5ths rule. Adverse impact is said to exist when the hiring rate of minority group members is less than 80% of the hiring rate of majority group members:

Adverse Impact = (# minority selected / # minority tested) / (# majority selected / # majority tested)

However, when there are small numbers of minority candidates (e.g., Native American), even small changes in hiring rates can produce dramatic differences in adverse impact ratios. There are other approaches that have strengths and weaknesses (e.g., 2 standard deviation rule, statistical significance, etc.). For example, relying on statistical significance is problematic when sample sizes are in the hundreds of thousands and even trivial effect sizes are statistically significant.

Murphy and Jacobs (2012) proposed a two-step regression-based approach for evaluating whether subgroup differences are meaningful versus ignorable (note it is assumed there is a minimally appropriate sample size and thus minimally acceptable statistical power):

- Regress the criterion on the KSAO predictors and the subgroup category (e.g., sex: female = 1 and male = 0).
- Interpret the effect size for the subgroup category and the statistical significance of the model R^2. Assuming the model R^2 is statistically significant, then consider the following:
  - If the model R^2 is equal to or less than .01, then the effect of the subgroup is trivial and there is no meaningful adverse impact.
  - If the model R^2 is greater than .01, then the effect of the subgroup is meaningful and there is possible evidence of adverse impact.

Note that it is important to first evaluate potential bias in the criterion scores. KSAO predictors get their importance from their relationships with criteria. A biased criterion will produce a biased selection and classification model. Following best practices in performance assessment, as noted above in Section 1.3.4, will help reduce the likelihood of criterion bias. Further, estimates of subgroup differences can be influenced by artifacts like unreliability and range restriction and should be interpreted (and possibly corrected) accordingly.

1.7.4. Differential Prediction

It is common to use moderated multiple regression to test for differential prediction, which occurs when different subgroups have different regression slopes and thus the KSAO predictor scores are more valid for one subgroup than another subgroup. Differential prediction means that even if members from different subgroups had the identical KSAO score, the criterion score will differ, and thus indicates some form of predictive bias (Berry, 2015).

It is important to consider sample size, reliability, range restriction, and statistical power when testing for differential prediction. Assuming these artifacts are not influencing the model estimation, the moderated regression model such as that shown below can be used. Here we assume there is only one predictor, cognitive ability, and we are testing for differential prediction due to sex.

$$\hat{Y} = b_0 + b_1(\text{COGABIL}) + b_2(\text{SEX}) + b_3(\text{COGABIL*SEX})$$

If the effect size for sex (b_2) is meaningful and statistically significant, and the change in model R^2 is greater than .01, there is evidence for potential sex bias in terms of intercept mean differences on the criterion. If the effect size for the moderator term (b_3) is meaningful and statistically significant, and the change in model R^2 is greater than .01, there is evidence for potential sex bias in terms of slope differences (predictive bias). This means that the slopes (validity) between the KSAO predictor (cognitive ability) and the criterion scores differ between men and women.

In general, there is some empirical evidence finding modest differential prediction for cognitively loaded KSAO predictors, such that prediction is lower for non-Whites, but there remain many questions and alternative explanations (see Berry, 2015 for a thorough review; also

see Aguinis, Culpepper, & Pierce, 2010; Aguinis & Smith, 2007; Sackett et al., 2017). Differential prediction for other types of KSAO assessments, such as personality, appear negligible, although more research is needed. Thus, it is critical to always test for differential validity in selection and classification models.

1.7.5. Ways to Balance Diversity and Validity

Because there are subgroup mean score differences on many KSAO assessments, and possibly differences in validity as well, it is critical to understand the consequences of different selection and classification models on diversity and validity. Diversity and validity are thus not synonymous, and achieving one goal (e.g., validity) may come at the expense of the other goal (e.g., diversity). The extent to which diversity or validity will be maximized is dependent on many factors, including:

- KSAO predictor factors
  - Validity
  - Intercorrelations
  - Subgroup mean score differences
  - Differential prediction or validity
- Criterion factors
  - Subgroup mean score differences
  - Nature of criterion dimension (e.g., task versus interpersonal)
- System factors
  - Weighting of KSAO predictors
  - Use of single versus multistage system
  - Compensatory versus noncompensatory system
  - Selection ratio
- Number of members in each subgroup
  - Sample size, range restriction, reliability of scores, and related artifacts

Because there are so many specific issues that can influence the results of any given selection and classification model, it is impossible to give precise guidance. Therefore, the following suggestions are offered:

- First, understand the nature of the KSAOs and why subgroup differences may theoretically exist (e.g., Cottrell, Newman, & Roisman, 2015).
- Second, review the many suggestions for balancing diversity and validity (e.g., Ployhart & Holtz, 2008; Sackett, Schmitt, Ellingson, & Kabin, 2001). These are summarized in Table 10. For example, there is sometimes a desire to create bands of scores that are psychometrically indistinguishable from each other. Selecting within these bands, but giving preference for minority candidates, can be a means to increase diversity. However, giving explicit preference to minority candidates is controversial and raise a number of practical and legal issues that need to be carefully considered (see Campion, Outtz, Zedeck, Schmidt, Kehoe, Murphy, and Guion, 2001).
- Generally speaking, measuring the full range of cognitive and noncognitive KSAOs needed to perform a job will help increase both diversity and validity.

Table 10. Suggestions for Reducing Subgroup Differences and Increasing Validity (adapted from Ployhart & Holtz, 2008)

TABLE 2

Strategies for Reducing Racioethnic and Sex Subgroup Differences and Adverse Impact

Strategy and premise	Effects on reducing subgroup differences	Comments	Key references
Category I. Strategies that use predictors with smaller subgroup differences than cognitive ability			
1. *Use alternative predictor measurement methods (e.g., interviews, work samples, assessment centers, situational judgment tests, biodata).* Using alternative predictor measurement methods will reduce subgroup differences because they measure multiple cognitive and non-cognitive KSAOs, frequently minimize reading requirements, may engender more favorable reactions, and/or are based on job performance tasks for which subgroup differences are smaller.	• Generally effective, but specific reductions are quite variable (Table 1).	• Predictors with smaller cognitive loadings produce smaller differences. • Differences tend to be larger in applicant than incumbent settings. • Magnitudes of reductions are affected by predictor type and subgroup. • Some methods decrease differences for one group but increase them for another. • Validity may be lower than for overall cognitive ability (sometimes marginally). • Lower reliability of alternative predictor measurement methods may attenuate subgroup differences and give the erroneous impression that they are much smaller. • Developing/administering/scoring alternative predictor measurement methods is expensive/time consuming. • Applicant faking may be an issue.	Hough et al. (2001); Schmitt et al. (1996)
2. *Use educational attainment or GPA as a proxy for cognitive ability.* Because GPA and/or educational attainment are related to conscientiousness and motivational constructs, in addition to cognitive ability, using them as proxies for cognitive ability will reduce adverse impact.	• Small to moderate reduction in subgroup differences compared to cognitive ability.	• Subgroup differences increase as educational attainment increases. • Less valid than cognitive ability. • Educational attainment *may* be more useful than GPA. • Research primarily for White–Black differences. • Applicant faking may be an issue when using self-reports.	Berry, Gruys, & Sackett (2006), Roth and Bobko (2000)

TABLE 2 (continued)

Strategy and premise	Effects on reducing subgroup differences	Comments	Key references
3. *Use specific measures of ability.* Specific (narrow) measures of cognitive ability (e.g., verbal, quantitative) have smaller subgroup differences than overall cognitive ability.	• Small to moderate reduction in racioethnic *d*-values compared to overall ability measures. • Male/female differences may be larger than for overall ability and may favor men (quantitative ability) or women (verbal ability).	• With broad measures of performance, validity may be lower than when using overall ability. • Generally useful only with specific criteria (e.g., reading proficiency). • May need to administer more predictors, potentially increasing costs and time for administration and scoring (and if administering several measures of specific abilities, the predictor battery may consequently assesses overall cognitive ability).	Hough et al. (2001)
Category II. Strategies that combine and manipulate scores			
4. *Assess the full range of KSAOs.* If cognitive ability is one of the most valid predictors but also exhibits the highest subgroup differences, then adding noncognitive predictors that are related to performance but engender smaller subgroup differences may reduce the overall subgroup difference of the predictor battery.	• Generally effective, but the magnitude of reduction depends on predictor validities and intercorrelations.	• Diminishing returns after adding four or more predictors. • The predictor with the highest validity will most determine the composite subgroup difference (when using regression-based weights). • Including a full battery of predictors usually produces higher validity. • Including more predictor KSAOs is expensive/time consuming. • Applicant faking may be an issue. • Must consider issues involved with adding and combining predictors in a battery (e.g., relative importance, incremental importance, incremental validity, and so on).	Bobko et al. (1999), LeBreton, Griepentrog, Hargis, Oswald, and Ployhart (in press), Ryan, Ployhart, and Friedel (1998); Sackett and Ellingson (1997), Sackett and Roth (1996), Schmitt et al. (1996)

TABLE 2 (continued)

Strategy and premise	Effects on reducing subgroup differences	Comments	Key references
5. *Banding and score adjustments.* There is no perfectly reliable predictor; acknowledging this unreliability by creating "bands" from within which scores cannot be empirically distinguished may increase racioethnic minority or female hiring.	• Reductions can be sizeable if using racioethnic minority or female preference within bands; otherwise reductions are small or nonexistent.	• Many factors influence effects of banding, including selection ratio, proportion of racioethnic minority or female applicants, and procedure for hiring within bands. • Racioethnic minority or female preference is usually illegal. • Recent questions about appropriate form of reliability estimate. • May lower validity.	Aguinis (2004), Campion et al. (2001), Murphy, Osten, and Myors (1995), Sackett and Roth (1991)
6. *Explicit predictor weighting.* Rather than simply summing the predictors or using regression-based weights, one may rationally give more weight to predictors with less adverse impact.	• Small to moderate reduction in subgroup differences.	• Greater reduction likely comes from choosing which predictors to put in the battery, rather than differential weighting within the battery. • Validity may be lowered with rationally derived weights. • Applicant faking may be an issue if using non cognitive predictors.	DeCorte (1999), DeCorte and Lievens (2003), Ryan et al. (1998)

TABLE 2 (continued)

Strategy and premise	Effects on reducing subgroup differences	Comments	Key references
7. *Criterion weighting.* Task and/or technical dimensions of performance are more strongly predicted by cognitive ability, whereas contextual/non technical dimensions of performance are more strongly predicted by non cognitive measures like personality. Emphasizing contextual/non technical dimensions will therefore reduce adverse impact through increasing the importance (validity) of non cognitive predictors.	• Small to moderate reduction in adverse impact when weighting contextual or non technical performance.	• Reductions are frequently not large unless selection ratio is high. • Criterion weighting will strongly influence validity, so cannot simply overweight contextual dimensions. • Applicant faking may be an issue if using non cognitive predictors.	Hattrup, Rock, and Scalia (1997), Murphy and Shiarella (1997)
Category III. Strategies that reduce construct irrelevant variance from predictor scores			
8. *Minimize verbal ability requirements to the extent supported by job analysis.* By assessing verbal ability only to the level supported by a job analysis, and/or using video-based predictors, this strategy reduces variance from subgroup differences in verbal ability and may enhance applicant reactions.	• Generally effective but the magnitude is variable.	• Must demonstrate equivalence when developing "lower verbal ability" alternative. • Must ensure verbal ability is not contributing to the validity of the alternative. • Developing/administering/scoring video-based or non written predictor methods is expensive/time consuming. • Verbal ability requirements cannot be lower than the minimum level identified in the job analysis.	Arthur, Edwards, & Barrett (2002), Sacco, Scheu, Ryan, Schmitt, Schmidt, and Rogg (2000), Sackett et al. (2001)

TABLE 2 (continued)

Strategy and premise	Effects on reducing subgroup differences	Comments	Key references
9. *Use "content free" items* that are not more (un)familiar to, or do not serve to advantage, any particular cultural subgroup.	• Small and inconsistent.	• Difficult to write items truly representative of cultures, but equivalent for all cultures. • Little theoretical support in selection contexts.	DeShon, Smith, Chan, and Schmitt (1998), Whitney and Schmitt (1997)
10. *Differential Item Functioning (DIF).* Removing items that demonstrate DIF will reduce subgroup differences.	• Small and inconsistent.	• DIF will favor both White (or male) groups and racioethnic minority (or female) groups for different items. • Very little success in developing theories or explanations of DIF—very empirically driven.	Hough et al. (2001), Sackett et al. (2001)
11. *Sensitivity review panels.* Developers of predictors use what is called a "sensitivity review panel" to examine items and ensure they are appropriate and non-offensive to all relevant subgroups.	• No data on effectiveness.	• They are frequently used in practice, but there is no strong empirical evidence supporting their use. • At present, main benefit is probably for public relations. • May be costly and time consuming to implement sensitivity review panels.	Reckase (1996)
12. *No time limits.*	• No clear reductions.	• White and racioethnic minority groups both improve; White group sometimes improves more. • Extending time limits may be expensive/time consuming.	Sackett et al. (2001)

Reprinted with permission from publisher.

1.8 Generalizing from Experimental to Operational Use

The operational use of selection and classification models is also influenced by a number of factors. There are many good descriptions of operational issues to consider (e.g., SIOP Principles, 2018; Van Iddekinge & Ployhart, 2008). Three issues are particularly important: cross-validity, developing local norms, and retesting.

1.8.1. Cross-Validity and Cross-Validation

When conducting research on new or experimental selection and classification models, it must be understood that the results (i.e., validity, subgroup differences, etc.) observed in the experimental setting will differ to some degree from the operational setting. Note an experimental setting can be with applicants or incumbents; the key feature of experimental settings is that the selection and classification model is not being used to make operational decisions. Likewise, conducting validation research using job incumbents (concurrent validation) may differ from the results observed when the selection and classification model is conducted on candidates (predictive validation). The reasons for any such differences include:

- Sample size
- Sampling variability
- Range restriction
- Unreliability
- Ratio of sample size to number of KSAO predictors
- Candidate response (e.g., knowledge of job) and motivational differences (e.g., faking)

These factors and artifacts may contribute to differences in effect sizes between experimental and operational settings. The most important factors are sample size and ratio of sample size to number of KSAO predictors (Schmitt & Ployhart, 1999). When differences are found in the effect sizes between experimental and operational settings, the effect sizes ae usually smaller in the operational setting—and hence is known as *shrinkage*. Shrinkage is especially a concern when regression or non-unit weights (or rational weights) are used (although sample-specific variability will still be present).

It is therefore important to always cross-validate KSAO predictor scores and relationships with criteria. More specifically, to estimate the population cross-validity if one re-estimated the regression equation infinitely in new samples (Raju, Bilgic, Edwards, & Fleer, 1997). Cross-validation seeks to estimate the effect size that will be found in the operational setting. There are two approaches to cross-validation.

- Formula-based estimates are generally preferable because they estimate validity on the entire sample and apply a formula to estimate the cross-validity (and amount of shrinkage) into the operational setting. Formula-based estimates can be used on experimental or operational samples. Raju et al. (1997) and Schmitt and Ployhart (1999) review a variety of different estimates. Although there is no single best estimate, one attributed to Burket (Formula 7 in Schmitt & Ployhart, 1999) tends to be appropriate in most situations for estimating the population cross-validity: $\rho_c = (NR^2 - k)/(R(N - k))$;

where R is the sample multiple correlation, N = sample size, and k = number of KSAO predictors. Formula 2 in Schmitt and Ployhart (1999) is also useful in more limited situations.

- Empirical cross-validation is used when there are very large sample sizes. One first splits the sample into 2/3 and 1/3. The larger 2/3 sample is used to estimate the weights through regression. These weights are then applied to the smaller 1/3 sample (called the holdout sample). The difference between the two results in terms of model R^2 is the amount of shrinkage. The model R^2 is the expected estimate for the operational setting, and the weights from the holdout sample are the weights to be used operationally. The empirical cross-validation approach has some serious limitations. First, it requires reducing the sample size, which increases standard errors and can affect the stability of regression weights. Second, the selection and classification model is still potentially affected by factors that affect the experimental testing session.

There is no reason one cannot use both formula and empirical approaches to cross-validation. However, the estimates form the formula-based approach are generally to be preferred (Guion, 2011).

1.8.2. Development of Local Norms

The USAF engages in large-scale selection and placement. Hence, it is possible to create local norms for different KSAO scores, overall and broken down by KSAOs and KSAO assessment methods, demographic subgroups, jobs and occupations, locations, and so on. Creating these local norms is extremely important for many purposes. First, they can enable estimates of unrestricted means and variances that are important for estimating range restriction. Second, they can inform the setting of cut scores, qualifying rates and levels, and related noncompensatory decisions. Third, they help establish estimates of subgroup demographic score differences.

1.8.3. Retesting

When candidates score below a cutoff on KSAO assessments, a natural question is whether they should be allowed to retest. Research to date suggests that organizations and applicants should consider retesting. Scores most often increase with retesting, but the increase can be due to job-related and job-unrelated reasons. Therefore, it is important to understand the factors that may affect retesting in any particular situation. For example, if retesting leads to higher scores on assessments that measure relatively stable KSAOs (e.g., cognitive ability), it suggests the score change has more to do with construct irrelevant variance (e.g., test familiarity). The criterion-related validity of retesting appears similar to the initial testing effort. However, sometimes retesting can produce lower scores or lower quality psychometric characteristics (e.g., lower reliability). See Van Iddekinge and Arnold (2017) for more details.

1.9 Saving Data and Reporting Results

It is very important to ensure the data are captured and results summarized so that one can understand what was done and how it was done.

- Saving data. Always save the data in the simplest file format possible that does not lose information. It is also important to have a data dictionary so that all variables are described and labeled, it is clear how variables were created, major judgment calls and decision points are explained, and so on. Analysis programs should be similarly saved and labeled. There should be sufficient detail so that someone completely unfamiliar with the project could understand the data and be able to conduct the analyses.
- Ensuring good data management and analysis practices is critical. For high-stakes selection and classification programs, it is good to have two sets of independent researchers (a) check the quality of the data and (b) conduct independent analyses (or spot check each other's analyses) using (c) different statistical software packages, as a means to enhance confidence in the results.
- Reporting results. Technical reports should conform to best practices in reporting and presenting validation findings. There are specific reporting guidelines provided in the SIOP Principles (2018), the APA Standards (2014), and the Office of Personnel Management (https://www.opm.gov/policy-data-oversight/assessment-and-selection/reference-materials/). Also see Jeanneret and Zedeck (2017), Guion (2011), and Van Iddekinge and Ployhart (2008).

1.10 Future Issues

The science and practice of selection and classification is evolving quickly. Such evolution is driven by technological, demographic, and societal changes. This final section briefly explores the technical and analytic topics and suggests AFPC/DSYX personnel stay current in their developments. Technology is rapidly changing the nature of psychological measurement and analysis in terms of data and analytics.

1.10.1. Big Data

Due to mobile devices and greater connectivity through digital systems, people now produce massive amounts of data. Big data is not simply a big dataset; big data refers to data that is high in volume, velocity, and variety. These three factors combine to create a new frontier in data and data analytics. Although regression may still be used on such data, regression models may not necessarily be best suited for big data. For example, as it becomes easier to combine data from multiple sources (e.g., organizational units, geographic areas), the data may be nested within the higher-level entities (e.g., performance scores differ across units due to unit differences in applying performance standards). When such nonindependence exists, the assumptions of regression are violated and the effects and/or significance tests may be biased. In such situations hierarchical linear models (also called random coefficient models) are preferable. Such models can account for the non-independence in scores, and therefore produce unbiased estimates for the predictors. These models are well-established and could offer more accurate estimates if the data are nested (see Pinheiro & Bates, 2000, for details). Therefore they are used frequently in validation of selection and classification systems.

1.10.2. New Analytics

Table 11 provides a listing of big data analytic models. There is much talk today about big data analytics, artificial intelligence, machine learning, algorithms, and so on. To date, there is little

published scientific evidence that supports the use of these analytics or shows clear benefits over existing approaches. However, there are growing concerns and legal challenges against such methods.

Table 11. Big Data Analytic Models

Big Data Models	Description
Ensemble	Employing a finite set of machine learning algorithms, essentially averaging across different learning algorithms to create a meaningful average prediction.
Decision Tree, Markov, Chain	May be used for regression (continuous data) or classification (nominal data); identify way to assign outcomes based on inputs.
Network (social, neural)	Methodology that examines links, notes, strengths of links, network centrality, and other relational features, to summarize the network.
Natural Language Processing and Text (mining, analysis, sentiment)	Algorithms intended to scan text and create scores capturing latent constructs discovered from the text.
Bayesian, Bootstrapping, Resampling	Resampling with or without replacement, as a means to provide more accurate estimates of effects and standard errors.
Machine Learning and Artificial Intelligence	A broad class of models that use a starting algorithm that then adapts based on new data and/or model fit.
Classification (random forest, cluster, factor)	Type of ensemble learning model that seeks to classify data, frequently using decision tree frameworks.
Spatial (association)	Models that focus on associations between physical locations (e.g., distance, geographic location).
Virtual Reality and Augmented Reality	Assessment methods that increase the visual and aural fidelity of the assessment experience.
Games and Gamification	Methods that seek to employ gaming principles to create assessments targeting specific KSAOs.

1.10.3. Big Data Recommendations

Big data approaches and analytics should be *explored*. There are several recent articles and chapters that explore big data approaches, and these provide a good introduction to the topic (e.g., Arthur, Doverspike, Kinney, & O'Connell, 2017; Behrend & Landers, 2019; Landers, Fink, & Collmus, 2017). The scientific merits of these methods will likely continue, and they will likely advance to provide new solutions for selection and classification models. However, at this point they should *not* be used operationally until there is sufficient evidence to support their use.

- The guidelines and principles discussed in this report still apply
- The technical, professional, legal, and ethical guidelines still apply
- For "big data" and "regular data," the question is which analytic technique:
  - best serves the purpose of the project
  - best fits the nature of the problem

- produces efficient and effective solutions

2.0 REFERENCES

Aguinis, H., Culpepper, S. A., & Pierce, C. A. (2010). Revival of test bias research in preemployment testing. *Journal of Applied Psychology, 95*, 648–680.

Aguinis, H., & O'Boyle, E. (2014). Star performers in twenty-first century organizations. *Personnel Psychology, 67*, 313-350.

Aguinis, H., & Smith, M. A. (2007). Understanding the impact of test validity and bias on selection errors and adverse impact in human resource selection. *Personnel Psychology, 60*, 165–199.

Aiken, J. R., & Hanges, P. J. (2017). Methods of combining assessments for employment decisions. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (2nd ed., pp. 388-396). NY: Routledge.

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

Arthur, W. Jr., Doverspike, D., Kinney, T. B., & O'Connell, M. (2017). Impact of emerging technologies on selection models and research: Mobile devices and gamification as exemplars. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (2nd ed., pp. 967-986). NY: Routledge.

Becker, T. E., Robertson, M., & Vandenberg, R. J. (2018). Nonlinear transformations in organizational research: Possible problems and potential solutions. *Organizational Research Methods, 22*, 831-866.

Behrend, T. S., & Landers, R. N. (2019). Introduction to the special issue on advanced technologies in assessment: A science-practice concern. *Personnel Assessment and Decisions, 5*, i-iii.

Berry, C. M. (2015). Differential validity and differential prediction of cognitive ability tests: Understanding test bias in the employment context. *Annual Review of Organizational Psychology and Organizational Behavior, 2*, 435-463.

Bobko, P., & Roth, P. L. (2010). An analysis of two methods for assessing and indexing adverse impact: A disconnect between the academic literature and some practice. In J. L. Outtz (Ed.) *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 29–49). NY: Routledge.

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.

Bommer, W. H., Johnson, J. L., Rich, G. A., Podsakoff, P. M., & MacKenzie, S. B. (1995). On the interchangeability of objective and subjective measures of employee performance: A meta-analysis. *Personnel Psychology, 48*, 587-605.

Brannick, M. T., Pearlman, K., & Sanchez, J. I. (2017). Work analysis. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (2nd ed., pp. 134-161). NY: Routledge.

Campbell, J. P., & Wiernik, B. M. (2015). The modeling and assessment of work performance. *Annual Review of Organizational Psychology and Organizational Behavior, 2,* 47–74.

Campion, M. A., Fink, A. A., Ruggeberg, B. J., Carr, L., Phillips, G. M., & Odman, R. B (2011). Doing competencies well: Best practices in competency modeling. *Personnel Psychology, 64*, 225-262.

Campion, M. A., Outtz, J. L., Zedeck, S., Schmidt, F. L., Kehoe, J. F., Murphy, K. R., & Guion, R. M. (2001). The controversy over score banding in personnel selection: Answers to 10 key questions. *Personnel Psychology, 54*, 149-185.

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

Cottrell J. M, Newman D. A, & Roisman, G. I. (2015). Explaining the black-white gap in cognitive test scores: Toward a theory of adverse impact. *Journal of Applied Psychology, 100(6)*:1713–1736.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243,* 1668-1674.

De Corte, W., Lievens, F., & Sackett, P. R. (2007). Combining predictors to achieve optimal trade-offs between selection quality and adverse impact. *Journal of Applied Psychology, 92,* 1380–1393.

Gatewood, R., Feild, H., & Barrick, M. (2011). *Human resource selection*. Mason, OH: Cengage Learning.

Guion, R. M. (2011). *Assessment, measurement, and prediction for personnel decisions* (2nd ed.). NY: Routledge.

Harvey, R. J. (1991). Job analysis. In M. D. Dunnette and L. M. Hough (Eds.), *Handbook of industrial and organizational psychology (pp. 71-163)*. Palo Alto, CA: Consulting Psychologists Press.

Hough L. M., Oswald F. L., & Ployhart, R. E. (2001) Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment, 9*, 152–194.

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

Jaeger, R. M. (1989). Certification of student competencies. In R. L. Linn (Ed.), *Educational Measurement* (3rd Edition) (pp. 485-514). NY: American Council on Education/ Macmillan.

Jeanneret, P. R., & Zedeck, S. (2017). Professional guidelines/standards. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (2nd ed., pp. 599-630). NY: Routledge.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.

Johnson, J. W. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate Behavioral Research*, 35, 1–19.

Johnson, J. W., & Oswald, F. L. (2017). Test administration and the use of test scores. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (2nd ed., pp. 182-204). NY: Routledge.

Kehoe, J. F., & Sackett, P. R. (2017). Validity considerations in the design and implementation of selection systems. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (2nd ed., pp. 56-92). NY: Routledge.

Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. *Journal of Applied Psychology, 98,* 1060–1072.

Landers, R. N., Fink, A. A., & Collmus, A. B. (2017). Using big data to enhance staffing: Vast untapped resources or tempting honeypot? In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (2nd ed., pp. 949-966). NY: Routledge.

Lievens, F., & Sackett, P. R. (2017). The effects of predictor method factors on selection outcomes: A modular approach to personnel selection procedures. *Journal of Applied Psychology, 102*, 43-66.

Locke, E. A. (1976). The nature and causes of job satisfaction. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology*, pp. 1297-1343. Chicago: Rand McNally.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

Morgeson, F. P., & Campion, M. A. (1997). Social and cognitive sources of potential inaccuracy in job analysis. *Journal of Applied Psychology, 82*, 627-655.

Murphy, K. R. (2009): Validity, validation and values. *The Academy of Management Annals, 3*, 421-461.

Murphy, K. R., & Jacobs, R. R. (2012). Using effect size measures to reform the determination of adverse impact in equal employment litigation. *Psychology, Public Policy, and Law, 18,* 477–499.

Murphy, K. R., & Myors, B. (1998). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests*. Mahwah, NJ: Lawrence Erlbaum.

O'Boyle, E., & Aguinis, H. (2012). The best and the rest: Revisiting the norm of normality of individual performance. *Personnel Psychology, 65*, 79-119.

Oswald, F. L., Putka, D. J., & Ock, J. (2015). Weight a minute…What you see in a weighted composite is probably not what you get! In C. E. Lance & R. J. Vandenberg (Eds.), *More statistical and methodological myths and urban legends* (pp. 187-205). NY: Routledge.

Pinheiro, J. C., & Bates, D. M. 2000. Linear mixed-effects models: Basic concepts and examples. In D. M. Bates & J. C. Pinheiro (Eds.), *Mixed-effects models in S and S-Plus*: 3-56. NY: Springer Verlag.

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

Ployhart, R. E., Schmitt, N., & Tippins, N. T. (2017). Solving the supreme problem: 100 years of recruitment and selection research at the *Journal of Applied Psychology*. *Journal of Applied Psychology, 102*, 291-304.

Ployhart, R. E., Schneider, B., & Schmitt, N. (2006). *Staffing organizations: Contemporary practice and research*. Mahwah, NJ: Erlbaum.

Ployhart, R. E., Weekley, J. A, & Dalzell, J. (2018). *Talent without borders: Global talent acquisition for competitive advantage*. Oxford University Press.

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

Rich, G. A., Bommer, W. H., MacKenzie, S. B., Podsakoff, P. M., & Johnson, J. L. (1999). Apples and apples or apples and oranges? A meta-analysis of objective and subjective measures of salesperson performance. *Journal of Personal Selling & Sales Management, 19*, 41-52.

Roth, P. L., Le, H., Oh, I.-S., Van Iddekinge, C. H., & Robbins, S. B. (2017). Who r u?: On the (in) accuracy of incumbent-based estimates of range restriction in criterion-related and differential validity research. *Journal of Applied Psychology, 102*, 802–828.

Sackett, P. R., Dahlke, J. A., Shewach, O. R., & Kuncel, N. R. (2017). Effects of predictor weighting methods on incremental validity. *Journal of Applied Psychology, 102*, 1421-1434.

Sackett, P. R., & Ellingson, J. E. (1997). The effects of forming multi-predictor composites on group differences and adverse impact. *Personnel Psychology, 50*, 707-721.

Sackett, P. R., Lievens, F., Van Iddekinge, C. H., & Kuncel, N. R. (2017). Individual differences and their measurement: A review of 100 years of research. *Journal of Applied Psychology, 102,* 254-273.

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative-action world. *American Psychologist, 56*, 302–318.

Sackett, P. R., & Wilk, S. L. (1994). Within-group norming and other forms of score adjustment in preemployment testing. *American Psychologist, 49*, 929-954.

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

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

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*(2), 262–274.

Schmitt, N., Arnold, J. D., & Nieminen, L. (2017). Validation strategies for primary studies. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (2nd ed., pp. 34-55). NY: Routledge.

Schmitt, N., & Chan, D. (1998). *Personnel selection: A theoretical approach*. Thousand Oaks, CA: Sage.

Schmitt, N., & Ployhart, R. E. (1999). Estimates onf cross-validity for stepwise-regression and with predictor selection. *Journal of Applied Psychology, 84,* 50-57.

Sellman, W. S., Russell, T. L., & Strickland, W. J. (2017). Selection and classification in the U.S. military. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (2nd ed., pp. 697-721). NY: Routledge.

Society for Industrial and Organizational Psychology. (2018). *Principles for the validation and use of personnel selection procedures*. Bowling Green, OH: American Psychological Association.

Smith, P. C. (1976). Behaviors, results, and organizational effectiveness: The problem of criteria. In M. D. Dunnette (Ed.), *Handbook of Industrial and Organizational Psychology*. Chicago: Rand McNally.

Smith, P. C. & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology, 47*, 149-155.

Song, Q. C., Wee, S., & Newman, D. A. (2017). Diversity shrinkage: Cross-validating pareto-optimal weights to enhance diversity via hiring practices. *Journal of Applied Psychology, 102*, 1636-1657.

Tippins, N. T., Solberg, E. C., & Singla, N. (2017). Decisions in the operational use of employee selection procedures. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (2nd ed., pp. 367-387). NY: Routledge.

Van Iddekinge, C. H., & Arnold, J. D. (2017). Retaking employment tests: What we know and what we still need to know. *Annual Review of Organizational Psychology and Organizational Behavior, 4*, 445-471.

Van Iddekinge, C. H., & Ployhart, R. E. (2008). Developments in the criterion-related validation of selection procedures: A critical review and recommendations for practice. *Personnel Psychology, 61*, 871-925.

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology, 90*, 108-131.

Woehr, D., & Huffcutt, A. (1994). Rater Training for Performance Appraisal. *Journal of Occupational and Organizational Psychology. 67*, 189-205. 10.1111/j.2044-8325.1994.tb00562.x.

APPENDIX SAMPLE CODE

```
/*USAF SELECTION & CLASSIFICATION CHAPTER SAS CODE*/
/*DR. ROBERT E. PLOYHART
/*11.26.2019*/


/*CODE TO TAKE A KNOWN CORRELATION MATRIX AND GENERATES THE RAW DATA*/
OPTIONS LS=80 PS=80;
PROC IML;
*************************************************************************
*****************;
************************ SET SIMULATION PARAMETERS
****************************************;
*************************************************************************
******************;

SEED1 = 3847;

N = 500; * NUMBER OF PEOPLE    ******************;
K = 5; * NUMBER OF VARIABLES ******************;

VC =
{1.00 .32 .17 .24 .08,
 .32 1.00 .25 .34 .10,
 .17 .25 1.00 .47 .11,
 .24 .34 .47 1.00 .53,
 .08 .10 .11 .53 1.00};

A=ROOT(VC); * CHOLESKEY DECOMPOSITION OF V ***********;

X= NORMAL(J(N,K,SEED1)); * GENERATE A MATRIX (X) OF RANDOM NORMAL VARIABLES;
X = X*A;
CREATE OUT FROM X; APPEND FROM X;
QUIT;


/* RENAMING THE VARIABLES */
DATA RAW1; SET OUT;
TRAINPERF=COL1 ;
COGABIL=COL2 ;
CONSCIENT=COL3 ;
ADAPT=COL4 ;
EMOTION=COL5 ;

RUN;

/* DESCRIPTIVES AND CORRELATION */
PROC UNIVARIATE PLOT; VAR TRAINPERF COGABIL CONSCIENT ADAPT EMOTION;RUN;
PROC CORR; VAR TRAINPERF COGABIL CONSCIENT ADAPT EMOTION;RUN;


/*BASELINE REGRESSION MODEL*/
PROC REG;
 MODEL TRAINPERF=COGABIL CONSCIENT ADAPT EMOTION/STB ;RUN;
```

```
 /*REDUCED REGRESSION MODEL*/
PROC REG;
 MODEL TRAINPERF=COGABIL CONSCIENT ADAPT /STB ;RUN;

 /*REDUCED REGRESSION MODEL*/
PROC REG;
 MODEL TRAINPERF=COGABIL CONSCIENT /STB ;RUN;
  /*REDUCED REGRESSION MODEL*/
PROC REG;
 MODEL TRAINPERF=COGABIL ADAPT /STB ;RUN;
   /*REDUCED REGRESSION MODEL*/
PROC REG;
 MODEL TRAINPERF=COGABIL /STB ;RUN;
```

LIST OF SYMBOLS, ABBREVIATIONS, AND ACRONYMS

Δ	Change
<	Less than
Σ	Sum
ρ	Corrected correlationA1 Air Staff
A1PT	Air Force Testing Policy
AF-WIN	Air Force Work Interest Navigator
AFPC/DSYX	Air Force Personnel Center/Strategic Research and Assessment
AFPC/DP3SP	Air Force Personnel Center Promotions, Evaluations, and Recognition Branch
AFHRL	Air Force Human Resources Laboratory
AFPD	Air Force Policy Directive
AFOQT	Armed Forces Officer Qualifying Test
AFRS	Air Force Recruiting Service
AFRS/RSOA	Air Force Recruiting Service/Operations Division Analysis Branch
AFSC	Air Force Specialty Code
AF-WIN	Air Force Work Interest Navigator
APA	American Psychological Association
ASVAB	Armed Services Vocational Aptitude Battery
ATST	Air Traffic Scenarios Test
BARS	Behaviorally-Anchored Rating Scale
b_0	Intercept; or the score on the criterion when the predictors are equal to zero
b_k	Regression weight associated with KSAO predictor X_k; how much of a change in Y is associated with a 1 unit change on the KSAO predictor
e	Error or residual term
EDPT	Electronic Data Processing Test
d	Standardized mean score difference
DoD	Department of Defense
GPA	Grade Point Average
HR	Human Resources
K	Number of KSAO predictors
KSAO	Knowledge, Skills, Abilities, and Other Characteristics
MCP	Minimally Competent Person
MTT	Multi-Tasking Test
N	Sample Size
O*NET	Occupational Information Network
OPM	Office of Personnel Management
p	Probability
ρ_c	The population cross-validity
PCSM	Pilot Candidate Selection Method
r_{xy}	Observed correlation between the KSAO predictor (X) and the criterion (Y)
r_{xx}	Reliability of the KSAO predictor score
r_{yy}	Reliability of the criterion score

R	Sample multiple correlation
R^2	Multiple correlation squared
RPA	Remotely-Piloted Aircraft
SAS	A statistical software suite for data management, advanced analytics, multivariate analysis, business intelligence, criminal investigation, and predictive analytics.
SD	Standard Deviation
SDI	Self Descriptive Inventory
SIOP	Society for Industrial and Organizational Psychology
SME	subject matter expert
TAPAS	Tailored Adaptive Personality Assessment System
TBAS	Test of Basic Aviation Skills
USAF	United States Air Force
Y	Criterion score
$\hat{Y}$	The predicted value for the criterion,