AFRL-RH-WP-TR-2021-0002

AIR FORCE PERSONNEL CENTER BEST PRACTICES GUIDE: TEST DEVELOPMENT AND VALIDATION

James M. LeBreton

Personnel Decisions Research Institutes, LLC

January 2021

Interim Report

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

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the Air Force Research Laboratory Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RH-WP-TR-2021-0002 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

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE (DD-MM-YY)	2. REPORT TYPE	3. DATES COVERED (From - To)
08-01-21	Interim	14-03-19 – 31-12-20

4. TITLE AND SUBTITLE

Air Force Personnel Center Best Practices Guide: Test Development and Validation

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

5b. GRANT NUMBER
Not applicable

5c. PROGRAM ELEMENT NUMBER
62202F

6. AUTHOR(S)
James M. LeBreton

5d. PROJECT NUMBER
5329

5e. TASK NUMBER
09

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

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

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

8. PERFORMING ORGANIZATION REPORT NUMBER

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

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

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

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

11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)
AFRL-RH-WP-TR-2021-0002

12. DISTRIBUTION/AVAILABILITY STATEMENT

Distribution Statement A: Approved for public released

13. SUPPLEMENTARY NOTES

Report contains color. AFRL-2021-0247, cleared on 2 February 2021

14. ABSTRACT

This series of reports consolidates the experience, wisdom, and tools the Air Force has accumulated in its selection and classification work, and blends them with best practice recommendations from industry. This entry covers test development and validation, beginning with an introduction to AFPC/DSYX and providing recommendations and best practices around test development and validation in the Air Force. 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.This report addresses five major topics: The first is validity and the validation process. The second covers test development and validation process. The third discusses using classical test theory to evaluate items and build tests. The fourth goes over using classical test theory to evaluate items and build tests. And the last section discusses item bias and test bias. The report also includes appendices containing annotated R code for conducting analyses described in the text proper.

15. SUBJECT TERMS

Selection, Validity, Validation, Testing, DSYX, Air Force,

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	114	19b. TELEPHONE NUMBER (Include Area Code) N/A

TABLE OF CONTENTS

List of Figures

List of Tables

FOREWORD

This report is one of a series of 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 (I/O) Psychology to provide guides to cover a variety of topics. Each begins with a section describing AFPC/DSYX and the background of their research to provide context for the series. This report addresses best practices on test development and validation, with an emphasis on I/O psychology, industry, and government.

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 the Air Force Personnel Center Strategic Research and Assessment branch (AFPC/DSYX) continues to develop assessments and refine selection and classification models for a large number of Air Force career fields.

We begin with an introduction to AFPC/DSYX. The background and history are covered, describing how the Air Force Human Resources Laboratory and its elimination left a need for providing research in human capital management. That was resolved in 2010 with funding to create 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 AFPC/DSYX can build on its capabilities.

The body of this report provides recommendations and best practices around test development and validation 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.

This report addresses a broad range of topics and is divided into five major sections. The first addresses validity and the validation process. In addition to defining validity, it reviews the current state of the science regarding the appropriate sources for accumulating validity evidence. The next section discusses steps in the test development and validation process. This section maps five levels of validation discussed in the Air Force Examining Activities Overview - Fiscal Year (FY) 2010-2011 onto the primary steps and activities suggested by a number of contemporary psychometricians.

Then, the report discusses using classical test theory to evaluate items and build tests, providing a detailed summary of the statistics used as part of item analysis and test evaluation under classical test theory. This section also introduces newer recommendations related to the estimation of internal consistency reliability that capture temporal stability at both the item-level and test-level. Next, it goes over using classical test theory to evaluate items and build tests, reviewing models applicable to dichotomously scored items and discusses the interpretation of item parameters, item information, the standard error of the estimate, building tests, and evaluating model fit. Finally, in the last section, the report discusses item bias and test bias, including the definition of bias and approaches for testing bias under classical test theory and item response theory. The report also includes appendices containing annotated R code for conducting analyses described in the text proper.

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 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 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

DSYX makes contributions to the Air Staff by following the mission as tasked by Air Force Manual (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 Air Force Base (AFB).

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

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

The DSYX Testing Toolbox

General Ability/Aptitude Tests

Air Force Officer Qualifying Test (AFOQT). The AFOQT is used to help select candidates for officer commissioning programs and to classify commissioned officers into utilization specialties such as manned aircraft pilot, RPA pilot combat system operators, air battle manager, or technical. 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 AFSCs. The second version is for use closer to commissioning when finalizing the AFSC assigned to a cadet upon commissioning.

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

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 DSYX Expertise and Resources Toolbox

Staff Expertise

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

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

Contractor Expertise

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

Forward Looking: The Future of 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 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 TEST DEVELOPMENT AND VALIDATION

1.1 Introduction

This report addresses a broad range of topics. Following the *Introduction,* the report is divided into five major sections. Section 1.2 addresses *Validity and the Validation Process.* After defining validity, it reviews the current state of the science regarding the appropriate sources for accumulating validity evidence. Section 1.3 discusses the *Steps/Stages in the Test Development and Validation Process.* This section maps the five levels of validation discussed in the Air Force Examining Activities Overview-FY 2010-20011 (pp. 25-26) onto the primary steps and activities suggested by a number of contemporary psychometricians. Section 1.4 discusses *Using Classical Test Theory to Evaluate Items and Build Tests.* It provides a detailed summary of the statistics used as part of item analysis and test evaluation under classical test theory. In addition to reviewing traditional recommendations, this section also introduces newer recommendations related to the estimation of internal consistency reliability using coefficient Omega (and its variants) and test-retest reliability using newer statistics that capture temporal stability at both the item-level and test-level. Section 1.5 covers topics related to *Using Classical Test Theory to Evaluate Items and Build Tests.* The section reviews models applicable to dichotomously scored items and discusses the interpretation of item parameters, item information, the standard error of the estimate, building tests, and evaluating model fit. It concludes with Section 1.6, which discusses *Item Bias and Test Bias.* Bias is defined and approaches for testing bias under classical test theory (CTT) and item response theory (IRT) are discussed. *Specific recommendations are set-off using bullet points.* Finally, appendices are included containing annotated R code for conducting analyses described in the text proper.

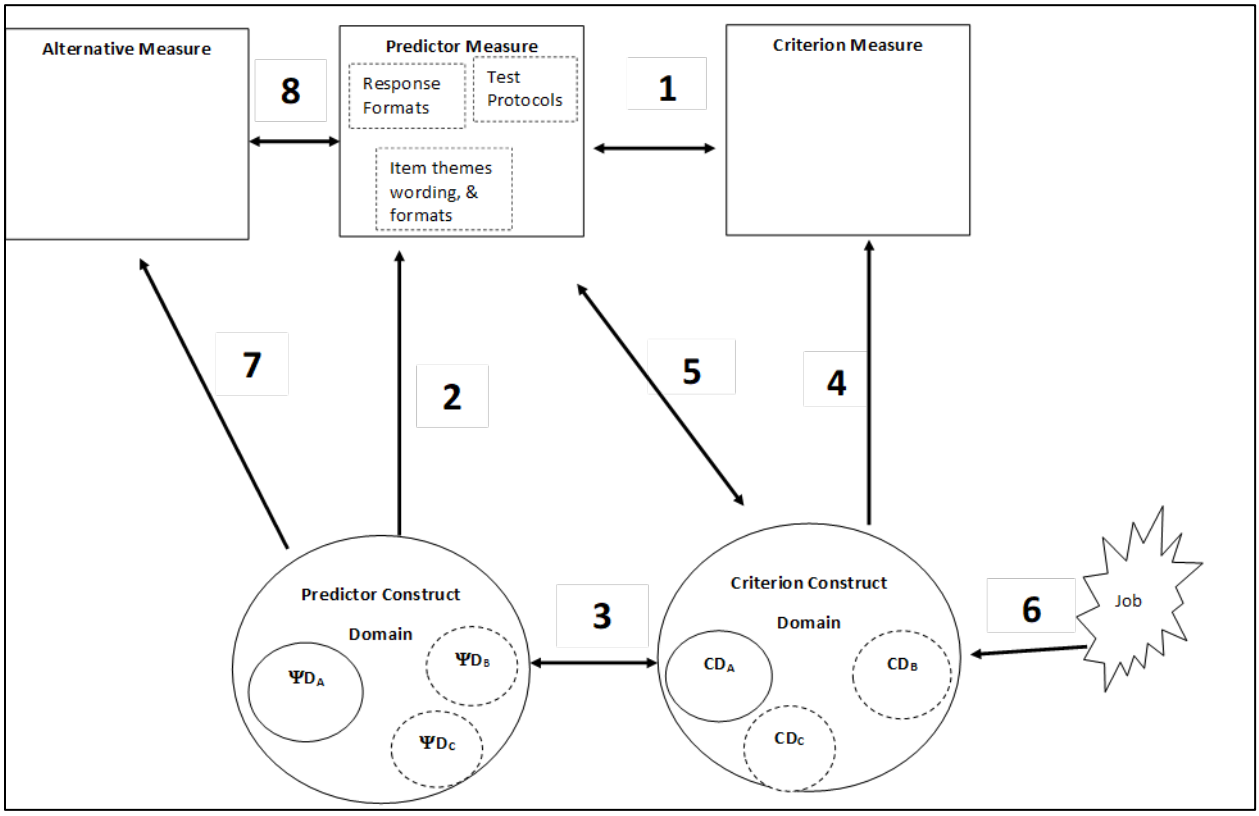
1.2 Validity and the Validation Process

Validity may be defined as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (p. 11; American Educational Research Association (AERA), et al., 2014). The *Principles for the Validation and Use of Personnel Selection Procedures* (2018) stated "Validity is the most important consideration in developing and evaluating selection procedures. Because validation involves the accumulation of evidence to provide a sound scientific basis for the proposed score interpretations, it is the interpretations of these scores required by the proposed uses that are evaluated, not the selection procedure itself" (p. 5). Thus, tests are not said to be valid or invalid, but rather the inferences drawn from test scores are said to be valid or invalid (e.g., a ruler may be used to draw valid inferences about a person's height, but invalid inferences about a person's weight).

1.2.1. A Conceptual Model for Discussing Validity and Validation

Figure 1 was derived from previous work by Binning and Barrett (1989) and Binning and LeBreton (2009) and introduces a framework for conceptualizing the process of accumulating validity evidence, specifically within the context of employee selection. Briefly, the development of any selection test should begin with a careful analysis of the target job. This job analysis helps to ensure that any selection test is "job relevant", per legal guidelines. The goal of the job analysis is to generate a job description (inference 6) by identifying the essential demands and requirements of a job and then translating those demands "into behavior-outcome units that

define the performance domain" (Binning & Barrett, 1989, p. 487). Once the performance domain has been clearly articulated, it can be used to develop selection tests (as well as criterion measures).



Note: Adapted from Binning & Lebreton (2009) Figure 1 (p. 289)

Figure 1. Framework for Conceptualizing Validity and Test Validation

One test development strategy is to identify the predictor constructs (e.g., verbal fluency; extroversion) that are hypothesized to be related to the criterion constructs (e.g., job performance; organizational withdrawal) that comprise the performance domain (inference 3). Once these predictor constructs have been identified, they are then used to guide the development of selection tests (e.g., ASVAB; see inference 2). An alternative test development strategy is to build selection tests (e.g., flight simulator test) that are believed to be more directly representative of the criterion constructs (inference 5). It is also possible to use the results of a job analysis to develop criterion measures (inference 4). Finally, it is possible to examine the usefulness of a selection test by using it to predict scores on criterion measures (inference 1). This is an admittedly simplified overview of this important framework for test validation. However, it will be referenced throughout the remainder of the report so some basic familiarity with the model will be helpful. The interested reader is directed to Binning and Barrett (1989) and Binning and LeBreton (2009).

1.2.2. Sources of Validity Evidence

There is a general consensus that validity should be considered a "unitary concept with different sources of evidence contributing to an understanding of the inferences that can be drawn from [test scores]" (p. 6; Principles, 2018; see also AERA et al., 2014; Binning & Barrett, 1989; Landy, 1986; Messick, 1995). Although different sources of validity evidence may be sampled using different strategies, the end goal of all personnel test validation efforts should be the demonstration that test scores predict relevant aspects of job performance/work behavior (Principles, 2018). Both the Standards and the Principles have identified five distinct sources of validity evidence: 1) relationships between test scores and other variables, 2) test content, 3) internal structure of the test, 4) response processes, and 5) consequences of testing. Of particular importance is the notation that no single source of validity evidence is to be considered "superior" to the other sources. Rather, the validity of inferences drawn from test scores are considered stronger when based on multiple, converging sources of evidence. Prior to reviewing these sources of validity evidence, it is important to first address the concepts of relevance, deficiency, and contamination in measurement, as those concepts are especially pertinent to the first two sources of validity evidence.

1.2.2.1 Measurement Relevance, Deficiency, and Contamination

Measures (both predictor and criterion) should include items/tasks that are representative of the underlying construct domain. This concept of relevance is important to researchers because it emphasizes the importance of sampling from the entire construct domain. If a measure (predictor or criterion) systematically fails to include items/tasks that are part of the construct domain, the measure is said to be deficient. In contrast, if a measure systematically includes irrelevant content (i.e., items/tasks that are unrelated to the construct domain), the measure is said to be contaminated. In contrast, when a measure (predictor or criterion) appropriately includes items/tasks that adequately sample from the breadth and depth of the focal construct domain (and appropriately excludes items/tasks asking about irrelevant constructs), then the measure is said to be relevant. When the target of measurement is a criterion construct, the discussion of relevance, deficiency, and contamination is sometimes denoted criterion relevance (Messick, 1995). These concepts are presented visually in Figure 2.

	Is included in the underlying construct domain	Is excluded from the underlying construct domain
Is included in the observed measure	RELEVANT	CONTAMINATION
Is excluded from the observed measure	DEFICIENT	CORRECT OMISSION

Figure 2. Measurement Relevance, Deficiency, and Contamination

1.2.2.2 Evidence Based on Relationships Between Test Scores and Other Variables

A critical component of any test validation process is to accumulate evidence supporting inferences linking test scores to other variables (i.e., measures of other constructs). Test scores may be related to a wide array of "other variables" including criterion measures (e.g., job performance; training performance; attrition), measures of the same (or similar) constructs, or measures of different constructs.

Convergent validity. Evidence of convergent validity is observed when scores on a focal test are highly correlated with scores on another test that purportedly measures the same (or a very similar) construct. More formally, "Convergent [validity] evidence exists when (a) test scores relate to scores on other tests of the same construct, (b) test scores from people who differ in the extent to which they possess the focal construct also differ in a predictable way, or (c) test score relate to scores on tests of other constructs that are theoretically expected to be related" (Binning & Barrett, 1989, p. 482). For example, two tests purportedly measuring verbal fluency should be highly correlated. Likewise, a test of verbal fluency is likely to be strongly correlated with a test of reading comprehension. For additional information regarding convergent validity, and the modeling of multiple sources of variance via confirmatory analysis, see Shaffer et al. (2016).

Divergent validity. In contrast, ***divergent (or discriminant) validity*** refers to the (lack of) correlation between scores on a focal test and scores on tests designed to measure psychologically distinct/different constructs. More formally, "Discriminant [validity] evidence occurs when test scores do not relate to scores on tests of theoretically independent constructs. Note that this discussion can apply equally to criterion measurement" (Binning & Barrett, 1989, p. 482). The second sentence in this quote is of particular importance-measures of criteria are measures of criterion constructs; thus, just as researchers should accumulate validity evidence for inferences drawn from test (predictor) scores, so too should they seek to accumulate validity evidence for inferences involving scores on criterion constructs. An example of accumulating evidence of divergent validity might take the form of testing the hypothesis that scores on a test of verbal fluency should be relatively uncorrelated with scores on tests designed to measure personality constructs (e.g., extroversion), job attitudes (e.g., job satisfaction), or job perceptions (e.g., justice climate).

Evidence of convergent and divergent validity may also be obtained by examining the pattern of relationships between test scores and measures of various demographic variables. For example, a researcher could predict that a measure of the motive for power (i.e., the desire or need to exert influence over social collectives and to take responsibility for the well-being of others; James, LeBreton, et al., 2013) would be positively related to military rank in large and diverse set of soldiers. In contrast, a researcher might predict that there would be no relationship between the motive for power and race/ethnic group membership.

Conceptually, evidence of convergent and divergent validity is represented in Figure 1 by inferences 2, 7, and 8. When the "alternative measure" refers to a different measure of the same (or highly similar) construct, these inferences are used to establish evidence of convergent validity. In contrast, when the "alternative measure" refers to a measure of a different construct, these inferences are then used to establish evidence of divergent validity. It is important to

remember that statistical tests of inference 8 (e.g., correlation coefficients; regression coefficients) are conditioned on inferences 7 and 2. These two inferences may be conceptualized as reflecting the psychological and psychometric fit between the latent construct(s) and the observed measures of those constructs.

Although bivariate correlations furnish an important initial test of convergent and divergent validity evidence, it is important to recognize that test scores may be subject to various forms of measurement error.

- *Researchers are encouraged to supplement bivariate correlational analyses with more sophisticated analyses that permit the more accurate decomposition and modeling of different sources of variability (see Shaffer, DeGeest, & Li, 2016; DeShon, 1998; Schmidt & Hunter, 1996).*

Criterion-related validity. Evidence of criterion-related validity is accumulated by showing that test scores are related to measures of one or more organizationally relevant criterion constructs. It is important to state explicitly, that "the criterion variable is a measure of some attribute or outcome that is operationally distinct from the test. Thus, the test is not a measure of a criterion, but rather is a measure hypothesized as a potential predictor of the targeted criterion (Standards, 2014, p. 17). Although measures of job performance are arguably the most commonly used criteria in applied psychology, criterion is a term used to refer to any organizationally valued outcome. Thus, relevant criteria might include: counterproductive workplace behaviors, organizational citizenship behaviors, attrition/turnover, intentions to quit or re-enlist, job satisfaction, job commitment, rate of promotion, and/or level of promotion (just to name a few alternative criteria). The accumulation of criterion-related validity evidence places a primary emphasis on inferences 4 and 1. Inference 4 represents the psychological (and psychometric) fit between the latent criterion construct and the observed indicator of that construct. Inference 1 represents the relationship between scores on predictor tests and scores on criterion measures.

- *Evidence for criterion-related validity (inference 1) may be accumulated by estimating the magnitude and significance of correlation or regression coefficients between criterion scores and predictor scores.*

According to the SIOP Principles, "A relevant, reliable, and uncontaminated criterion measure(s) is critically important" to any criterion-related validation study (p. 15). The most critical of these requirements is relevance-a criterion is said to be relevant when "it reflects the relative standing of employees with respect to some [organizationally valued outcome]" (p. 15). Arguably, the second most important requirement for a criterion measure is reliability (e.g., a psychometric index of inference 4). This is especially true when statistical tests of inference 1 are disattenuated for measurement error in criterion measures (e.g., inference 4). That is, the reliability of the criterion is important, especially if one is planning to "correct" correlations for measurement error. The correction equations are only as accurate as the point-estimates for reliability that are being plugged into those equations. Also, it is important to remember that underestimates of reliability (i.e., conservative estimates) will yield overestimates of validity (i.e., liberal estimates).

- *Researchers should base tests of criterion-related validity on criteria that are both job relevant and highly reliable.*

Some authors have raised concerns that the meta-analytic point estimates of criterion reliability based on supervisor ratings (r_{yy} = .52) and peer ratings (r_{yy} = .42; see Ones, Viswesvaran, & Schmitt, 1993) may be problematic (for the most recent discussion of these issues, the reader is encouraged to review the exchange in Volume 7, Issue 4 of *Industrial and Organizational Psychology: Perspectives on Science and Practice*). For example, some researchers have suggested that these point estimates may be underestimated because the traditional correlations used to estimate reliability may have been attenuated due to variance restriction on the criterion measure (LeBreton, Burgess, Kaiser, Atchley, & James, 2013), or may have been incorrectly computed due to ignoring important sources of variability in performance ratings (variance due to supervisors; variance due to subordinate by supervisor interactions; DeShon, 2003; Murphy & DeShon, 2000), or they may have been incorrectly computed because data may have been obtained using an ***ill-structured measurement design*** (Putka, Le, McCloy, & Diaz, 2008). Ill-structured measurement designs are common in the organizational sciences and represent a hybrid between a perfectly nested design (i.e., all subordinates are nested within a single leader who evaluates each subordinate) and a perfectly crossed design (i.e., all subordinates are rated by the exact same set of leaders). To better illustrate the differences between a nested, crossed, and ill-structured measurement design, consider a scenario where 10 job candidates are in an assessment center comprised of three different exercises. The pattern of ratings summarized in Table 1 reveals a crossed designed was used with Exercise #1, a nested design was used with Exercise #2, and an ill-structured design was used with Exercise #3.

- *Researchers should be mindful of the extent to which their data conform to a fully nested design, a fully crossed design, or an ill-structured measurement design (Putka et al., 2008) and use the appropriate equations for estimating reliability.*

Table 1. Illustration of Crossed, Nested, and Ill-structured Measurement Designs

Exercise #1: In-Basket (crossed-design)

Candidate	*Assessor 1*	*Assessor 2*	*Assessor 3*	*Assessor 4*	*Assessor 5*	*Assessor 6*
1	x	x	x	x	x	x
2	x	x	x	x	x	x
3	x	x	x	x	x	x
4	x	x	x	x	x	x
5	x	x	x	x	x	x
6	x	x	x	x	x	x
7	x	x	x	x	x	x
8	x	x	x	x	x	x
9	x	x	x	x	x	x

Exercise #2: Role-Play (nested design)

Candidate	*Assessor 1*	*Assessor 2*	*Assessor 3*	*Assessor 4*	*Assessor 5*	*Assessor 6*
1	x	x				
2	x	x				
3	x	x				
4			x	x		
5			x	x		
6			x	x		
7					x	x
8					x	x
9					x	x

Exercise #3: Leaderless Group Discussion (ill-structured design)

Candidate	*Assessor 1*	*Assessor 2*	*Assessor 3*	*Assessor 4*	*Assessor 5*	*Assessor 6*
1	x	x	x			
2	x		x	x		
3					x	x
4	x		x	x		
5	x	x				x
6		x	x			
7	x			x	x	
8	x	x				
9		x	x	x		

Predictive vs. concurrent validation designs. Criterion-related validation studies have traditionally been classified into two general categories based on when test scores and criterion measures are collected. When researchers adopt a predictive validation design, they commit to collecting criterion data after collecting data on the focal test (Standards, 2014). Thus, using this design, evidence of criterion-related validity exists when test scores predict subsequent (i.e., future) scores on the criterion. For example, scores on the ASVAB may be collected during

MEPS testing and scores on some criterion measure (e.g., job performance) may be collected 9 to 12 months downstream.

In contrast, when researchers adopt a ***concurrent validation design,*** they essentially collect predictor and criterion data at roughly the same time (Standards, 2014). Thus, evidence of criterion-related validity exists when there is a relationship between concurrently (or simultaneously) collected test scores and criterion scores. For example, researchers might collect data on the Remotely Piloted Aircraft (RPA) test from a sample of current remote pilots and at the same time obtain some type of performance data on those pilots (e.g., supervisory ratings of task performance; objective evaluations of multi-tasking effectiveness). In this instance, it is important to recognize that the variability in RPA test scores may be restricted, thus attenuating the correlation used to furnish evidence of criterion-related validity.

Corrections for statistical artifacts. Sample data will furnish imperfect estimates of population-level correlations and/or regression coefficients used to furnish evidence of criterion-related validity. The gap between observed sample estimates and the unobserved population estimates will fluctuate as a function of sampling error, measurement error (in both the predictor and criterion measures), and range restriction. Although equations exist that permit researchers to "correct" observed relationships for these statistical artifacts, it is important that any integration of these equations into practice is approached with great care and caution-especially if one is accumulating validity evidence using psychometric meta-analysis (Schmidt & Hunter, 2015). Several recent studies have documented that some of the important statistical assumptions that underlie psychometric meta-analysis may be untenable, or at least regularly violated (Köhler, Cortina, Kurtessis, & Gölz, 2015; Yuan, Morgeson, & LeBreton, 2020). Thus, if researchers wish to make corrections for statistical artifacts to observed correlations or regression coefficients, they are encouraged to do so judiciously.

- *When the variability in predictor (or criterion) scores is artificially restricted (i.e., a sample that is not fully representative of the target population), then researchers may correct observed correlations for range restriction.*

The reader is directed to Sackett and Yang (2000) and Schmidt, Oh, and Le (2006) for detailed guidance on matching range restriction corrections to different research scenarios; and, to Roth, Le, Oh, Van Iddekinge, and Robbins (2017) for a discussion on using applicant vs. incumbent samples when making corrections for range restriction.

Schmidt, Hunter, and Urry (1976) aptly noted, "In the typical validation study, the criterion reliability, as well as the test validity, is available only on the restricted group. Both coefficients should be corrected first for restriction of range" (p. 475). Unfortunately, the most commonly referenced reviews of criterion reliability failed to make such adjustments (cf., LeBreton et al., 2003; Viswesvaran, Ones, & Schmidt, 1996). Note also that the correction is agnostic with respect to whether values are estimated using the population of qualified candidates or the population of all candidates. The analyst must correctly align their analysis with the inference they wish to draw. There must be a theoretical rationale for the selection of the group being used to derive the values.

- *When researchers wish to correct observed correlations for measurement error in the criterion, they are strongly encouraged to correct reliability estimates for potential range restriction prior to using those reliability estimates in subsequent correction equations (LeBreton et al., 2003).*

Incremental validity. Evidence of incremental validity is accumulated by showing that the addition of an instrument improves our ability to predict some criterion of interest. At its most basic, "validity must be claimed for a test in terms of some increment in predictive efficiency" over other information already gathered (Sechrest, 1963, p. 154). When evaluating, if an assessment has incremental validity, the most common approach is to use multiple regression with the criterion of interest as the outcome. The first step in the regression model would include the initial assessment used (e.g. ASVAB completed during MEPS testing). The second step in the regression model would include the additional assessment (e.g., AFOQT). An inference concerning the incremental validity of a selection test would be supported by a significant change in the regression model R^2 after adding in the additional predictor (i.e., ΔR^2; Schmidt & Hunter, 1998). The "significance" of the change may be defined as statistical significance, practical significance (i.e., effect sizes), or both. It is important to note that when the two assessments used are less correlated with each other, there will be greater utility with the additional assessment. Recommendations related to testing for incremental validity are revisited later in the report.

1.2.2.3 Evidence Based on Test Content

Evidence of content-related validity is accumulated by documenting how the test content, defined as the "themes, wording, and format of the items, tasks, or questions on a test" (p. 14) are representative of the construct purportedly being measured by those items, tasks, or questions (Standards, 2014). In addition, test content may include the instructions, response formats, and various test administration protocols (Principles, 2018).

Binning and LeBreton (2014) clarified how content validity is relevant for the accumulation of evidence for tests developed to measure predictor constructs and tests that are designed to more directly measure criterion constructs. For example, a researcher may be interested in building tests designed to measure psychological (predictor) constructs such as cognitive ability, bravery, and cooperation (denoted $\psi_{DA}, \psi_{DB}, \psi_{DC}$ in Figure 1). In this context, content validity evidence should be accumulated to help support both inference 2 and inference 3. As Binning and LeBreton (2014) noted, "Inference 2 is supported by evidence that a given predictor adequately samples from a specific psychological [construct domain]" (p. 490).

- *When accumulating content validity evidence for selection tests designed to measure predictor constructs, researchers should evaluate the items/tasks to ensure they are representative of their respective construct domains (inferences 2 and 3).*

In contrast, when a predictor test has been designed to more directly assess aspects of the criterion construct domain, then accumulating evidence to support inference five becomes the focus of any content validation effort. Within the context of employee selection, evidence of content validity is accumulated by demonstrating (logically or empirically) that the items, tasks, or questions comprising the test are closely related to the actual work activities/tasks and work

outcomes comprising the criterion construct domain (Binning & Barrett, 1989; Binning & LeBreton, 2009; Principles, 2018). For example, if a job analysis results in the identification of criterion constructs including accuracy of data entry and customer service orientation (denoted C_DA and C_DC in Figure 1), then a researcher might build a selection test that directly measures behaviors identified as part of this construct domain (e.g., walk-through performance tests; assessment centers; other work simulations).

- *When researchers are accumulating content validity evidence for selection tests designed to measure criterion constructs, they should evaluate the items/tasks to ensure they are representative of the actual work activities/tasks/outcomes comprising the criterion construct domain.*

Statistical tools for evaluating evidence of content validity. Colquitt, Sabey, Rodell, and Hill (2019) distinguished between two aspects of content validity. Definitional correspondence refers to the "degree to which a scale's items correspond to the construct's definition" (p. 1243). Definitional distinctiveness refers to "the degree to which a scale's items correspond more to the focal construct's definition than to the definitions of other orbiting constructs" (p.1243). Essentially, definitional correspondence gets at the correct mapping of items onto constructs, whereas definitional distinctiveness is the ability of the items to discriminate/distinguish between the focal construct and other constructs. Colquitt et al. (2019) reported the results of a large study designed to evaluate two approaches for accumulating content validity evidence associated with both definitional correspondence and definitional distinctiveness.

Both approaches involve judges reviewing the definitions of multiple constructs and a set of items designed to measure these constructs. The judges, typically laypersons, not subject matter experts, are then tasked with sorting the items into the correct construct definition. The first approach, introduced by Anderson and Gerbing (1991), allows researchers to estimate two statistics. The proportion of substantive agreement (p_{sa}) provides an index of definitional correspondence and is computed by taking the number of judges who correctly match the item with the construct and by the total number of judges:

$$p_{sa} = \frac{number\ of\ correct\ judges}{number\ of\ total\ judges} \quad (1)$$

The p_{sa} index assumes values ranging from 0 to 1, with higher scores indicating greater degrees of definitional correspondence.

The substantive validity coefficient (c_{sv}) provides an index of definitional distinctiveness and is computed by taking the number of times the judges correctly match an item with the construct and subtracting the maximum number of times the item was incorrectly matched to any other construct. The difference is then divided by the total number of judges:

$$c_{sv} = \frac{(number\ of\ correct\ judges - maximum\ number\ of\ mismatches)}{total\ number\ of\ judges} \quad (2)$$

As Colquitt et al. noted, "The c_{sv} statistic ranges from -1 to 1, achieving the former value when no judges classify an item correctly and all do so incorrectly and the latter value when all judges classify an item correctly and none do so incorrectly" (p. 1244).

The second approach to estimating content validity that Colquitt et al. (2019) reviewed was derived from early research by Hinkin and Tracey (1999). This approach also relies on non-expert judges matching items with construct definitions. However, rather than sorting items into separate construct silos, the Hinkin and Tracey approach has judges rate the degree of item-construct correspondence using a Likert-type response scale. Building off this work, Colquitt et al. offered two content validity indexes paralleling the indexes of Anderson and Gerbing (1991). The first index is referred to as the Hinkin-Tracey correspondence index and is computed by taking the average judges rating and dividing by the number of points on the rating scale:

$$htc = \frac{average\ definitional\ correspondence\ rating}{number\ of\ points\ on\ the\ rating\ scale} \quad (3)$$

Thus, higher scores indicate greater definitional correspondence (e.g., when all judges select the maximum rating of correspondence, the htc willl be equal to 1).

Colquitt et al.'s (2019) second statistic was referred to as the Hinkin-Tracey distinctiveness index and is estimated by computing the signed differences between the correspondence rating for the intended construct and the correspondence rating for the orbiting constructs. These signed differences are averaged and then divided by the number of scale points minus 1. This statistic, denoted *htd*, "would have a positive value when items received higher ratings on the intended construct than on the orbiting constructs and a negative value when items received lower ratings on the intended construct than on the orbiting ones" (p. 1248):

$$htd = \frac{average\ of\ all(intended\ correspondence\ rating - orbiting\ construct\ rating)}{number\ of\ points\ on\ the\ rating\ scale - 1} \quad (4)$$

Colquitt et al. (2019) reported the results of a large content validation study that used both the Anderson and Gerbing (1991) and Hinkin and Tracey (1999) statistics. Data was collected from a total of 6,240 participants who evaluated subsets of 112 different scales. The authors provided descriptive statistics on these indexes of content validity and examined how various aspects of the scales were associated with these indexes (e.g., number of items, number of reverse coded items, magnitude of reliability coefficient, etc.). The authors concluded by providing a set of evaluative guidelines that could be used in future test development and validation studies (see especially their Table 5 on page 1257).

- *Researchers should compute and interpret indices of definitional correspondence and definitional distinctiveness (Colquitt et al., 2019) when seeking to accumulate validity evidence based on the content of the test.*

1.2.2.4 Evidence Based on the Internal Structure of the Test

Validity evidence may also be accumulated by examining the internal structure of a selection test. In doing so, researchers will likely examine the patterns of covariance between test items to

determine whether the pattern is consistent with the proposed constructs. For example, exploratory (or confirmatory) factor analyses may be used to explore (or confirm) the covariance structures in a set of items. However, as noted in the Principles (2018), "Inclusion of items in a selection procedure should be based primarily on their relevance to a construct or content domain and secondarily on their intercorrelations" (p.32). Essentially, different items based on different construct models will likely require different statistical analyses when seeking validity evidence based on the internal structure of the test. For example, when the construct model used to develop a test posits a single, unidimensional construct, researchers would focus on an analysis of item homogeneity and the presence of a single, dominant factor. When the construct model used to develop a test posits multiple dimensions or second order factors, then the analyses used to accumulate validity evidence would necessarily differ.

The Standards (2014) suggested that another way to accumulate evidence based on internal structure is to examine whether tests (or the items comprising them) are psychometrically equivalent or invariant across different groups of test takers (e.g., gender; race; age). These issues will be revisited later in the report when discussing tests of psychometric bias (e.g., differential item functioning/differential test functioning; measurement equivalence/invariance).

1.2.2.5 Evidence Based on Response Processes

Another form of validity evidence may be accumulated by verifying the processes individuals use when completing the test (Principles, 2018; Standards, 2014). This form of evidence is relevant for constructs that "involve more or less explicit assumptions about the cognitive processes engaged in by test takers" (Standards, 2014, p. 15). Evidence based on response processes may be collected directly from test takers by asking them about their response strategies or asking them to engage in a verbal protocol analysis as they complete the test. This approach may be particularly useful for tests designed to measure phenomena such as generating novel or creative solutions to problems, evaluating and weighting the quality and quantity of information prior to making a particular decision, and other cognitively loaded tasks. Alternatively, for tests measuring more overt behaviors (e.g., flight simulator), it may be possible to observe individuals as they complete the test. Other ways to accumulate evidence may include an examination of response times to computerized assessments or extracting information about the pattern and duration of visual attention using eye tracking software.

1.2.2.6 Evidence Based on the Consequences of Testing

Finally, validity evidence may also be provided by an examination of the consequences or outcomes (intended and unintended) that result from the use of a test. As noted in the Principles (2018), "Although evidence of negative consequences may influence policy or practice decisions concerning the use of predictors, the Principles and the Standards take the view that such evidence is relevant to inferences about validity only if the negative consequence can be attributed to the measurement properties of the selection procedure itself" (p. 8). Thus, using a test of physical strength would likely result (on average) in men receiving higher test scores than women, resulting in the unintended consequence of hiring many more men than women. This unintended consequence would only be relevant to the evaluation of validity if these group differences were driven by measurement properties of the test (i.e., psychometric bias) rather

than true, group mean differences. Strategies for identifying psychometrically biased items are discussed later in the chapter.

1.3 Steps/Stages in the Test Development and Validation Process

The development and validation of a personnel test for use in selection or classification is a multi-stage, iterative, process. Contemporary recommendations for scale development and validation closely mirror the five stages of test validation referred to by the Air Force as the Selection and Classification Test Acquisition Process. These five stages are summarized based on information contained in the Air Force Examining Activities Overview-FY10-11 (AFEAO, 2010) and are used as an organizing framework for integrating additional recommendations offered by other psychometric sources (see Table 2).

1.3.1. Level 1 Validation-Determination of Mission Need

"The initial phase of validation is determining whether there is a need or problem that requires more than a "quick fix" but could potentially be solved by a specific test or measure. At this level of validation the test or measure is developed or fine-tuned in an attempt to solve the identified problem or meet the need that was identified" (p. 25; AFEAO, 2010).

This stage of validation encompasses several sub-stages including: needs/job analysis, construct specification/definition, test development, and pilot testing.

1.3.1.1 Level 1a: Needs analysis/job analysis

The first step in test validation to determine the purpose of the test. In doing so, it will be important to determine what knowledge, skills, and abilities (KSAs) are necessary for the mission, so that a test can capture and measure those that are relevant. Conducting a thorough and comprehensive job analysis can tell you just that, since the goal of a job analysis is to "[discover, understand, and describe] what people do at work" (Brannick, Levine & Morgeson, 2007, p. 1). Whereas there are many methods of conducting a job analysis, the basic building blocks are as follows (see also Table 3):

- *Researchers should determine the preferred type or form of job data to collect.*

The first step before conducting a job analysis is to determine which descriptors (i.e. type of job data) to collect. This is largely dependent on the purpose of conducting the job analysis. For example, if the purpose is to identify those who might be successful on some mission, it might be important to collect descriptors that tell us about employee characteristics on the job (Brannick et al., 2007). These might include responsibilities (how much authority or accountability an employee has), personal job demands (physical demands), worker activities (focusing on what is going on inside of a worker's mind, how decisions are made, how problems are solved, etc.), work activities (observable behaviors performed on the job) or critical incidents (stories about on the job successes or failures).

Table 2. Summary of the Recommended Steps and Phases Associated with Test Development and Validation

Allen & Yen (1979)	Ghiselli, Campbell, & Zedeck (1981)	Crocker & Algina (1986)	Hambleton, Swaminathan, & Rogers (1991)	Hinkin (1998)	Air Force SCTAP (2010)	Standards (2014)
1. Plan the test	1. Defining a test	1. Identify the purpose of the test	1. Define Target Information Function	1. Item generation	Level 1: Determination of mission need	Phase 1: Test specifications
2. Write items	2. Specify test objectives	2. Identify behaviors representing the construct or	2. Iteratively select items to reproduce target curve	2. Questionnaire administration	Level 2: Concept exploration	Phase 2: Item development and review
3. Collect data	3. Item analyses	3. Prepare a set of test specifications	3. After adding each new item, estimate test information function.	3. Initial item reduction	Level 3: Program definition and risk reduction	Phase 3: Administration and scoring protocols
4. Item analysis	4. Empirical vs rationale keying	4. Construct initial item pool	4. Select items until test function approximates target function	4. Confirmatory factor analysis	Level 4: Engineering and manufacturing development	Phase 4: Test revisions
5. Finalize & norm the test	5. Weight items	5. Review & revise items		5. Convergent & divergent validity	Level 5: Production, deployment, operational support, and on-going monitoring	
	6. Cross-validate	6. Initial item tryouts		6. Replication		
	7. Advanced analysis (IRT)	7. Field-test items				
		8. Item analysis & item revision				
		9. Validities studies for the final test				
		10. Guidelines for test administration & use				

- *Researchers should determine the method or methods used for collecting the relevant job data.*

A wide array of data collection modalities or methods exist for collecting job analysis data. As we saw the descriptors depend on the purpose of the job analysis, so to does the degree to which one job analysis method is preferred over the other (e.g., develop criterion measures; develop predictor measures; develop compensation model; etc.) and context-specific factors related to the job and the organization. As noted in Table 3, the types of methods used to collect job analysis information may range from observing incumbents working in the job, to group interviews, to standardized questionnaires asking about job or worker characteristics, to retrieving information from the existing literature or archival sources, to job analysts actually performing the job (Brannick, et al., 2007; Cascio & Aguinis, 2019; Gatewood, Feild, & Barrick, 2015).

Table 3. Summary of Job Analysis Building Blocks

Descriptor	**Method of Data Collection**
1. Organization philosophy and structure 2. Licensing and other government-mandated requirements 3. Responsibilities 4. Professional standards 5. Job context 6. Products and services 7. Machines, tools, work aids, and checklists 8. Work performance indicators 9. Personal job demands 10. Elemental motions 11. Worker activities 12. Work activities 13. Worker trait requirements 14. Future changes 15. Critical incidents	1. Observing 2. Interviewing individuals 3. Interviewing groups 4. Technical conferences 5. Questionnaires 6. Diaries 7. Equipment-based methods 8. Reviewing records 9. Reviewing literature 10. Studying equipment design specifications 11. Doing the work
Sources of Job Analysis Data	**Units of Analysis**
1. Job analyst 2. Job holder's supervisor 3. High-level executive 4. Job holder 5. Technical expert 6. Organizational training specialist 7. Clients or customers 8. Other organizational units 9. Written documents (for example, records, equipment specifications) 10. Previous job analyses	1. Duties 2. Tasks 3. Activities 4. Elemental motions 5. Job dimensions 6. Worker characteristic requirements 7. Scales applied to units of work 8. Scales applied to worker characteristic requirements 9. Qualitative versus quantitative considerations

Note. Reproduced from "Job and work analysis: Methods, research, and applications for human resource management", Brannick, Levine, & Morgeson, 2007, Table 1.3, p. 19.

- *Researchers should determine the sources for obtaining relevant job information.*

Related to the previous point, it is important to ascertain where the job-relevant data will be obtained (e.g., from incumbents, supervisors, subordinates, trained observers, organizational archives; see Table 3 for summary). When the job analysis sources are human beings, it is important that researchers are mindful of the various social-cognitive biases that may distort the quality of the job analysis data. Table 4 was reproduced from Morgeson and Campion (1997) and provides a summary of how 16 different social-cognitive biases could impact the quality of job analysis information (for a detailed discussion, the reader is directed to Morgeson & Campion, 1997). Often, the source of data will be a function of the method of data collection (e.g., if you are interviewing job incumbents, your method is interviewing, your source is job incumbents).

- *Researchers should determine the unit of analysis for the job analysis.*

- *Researchers should seek to minimize the impact of social-cognitive biases when collecting data as part of a job analysis.*

    - *When possible, collect and score data using systematic and structured protocols.*

    - *When possible, collect and score data using multiple sources of information.*

Finally, one needs to determine the unit of analysis in which to report the data we have collected (i.e., how we report the work activities that we collected via interviews with incumbents). Like building blocks 1-3, this will be dependent on the choices we have made thus far. For example, one way to summarize, analyze and report the data on work activities is to break down the activities into their elemental motions. Another way to summarize work activity data might be to report the requirements needed to perform such work activities. Additionally, instead of just reporting the list of work activities gathered, job analysts may apply scales to these lists. For example, a common scale used is to ask how important an activity is and how frequently it is performed (Brannick et al., 2007). These scales can help determine how central each activity is to the job (usually those that are either very important, very frequent, or very important and frequent).

Table 4. Social and Cognitive Sources of Potential Inaccuracy and Their Hypothesized Effects on Job Analysis Data

	Likely Effect on Job Analysis Data					
Source of Inaccuracy	**Interrater Reliability**	**Interrater Agreement**	**Discriminability between Jobs**	**Dimensionality of Factor Structure**	**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	✓			✓		

Note. Reproduced from "Social and cognitive sources of potential inaccuracy in job analysis," Morgeson & Campion, 2005, *Journal of Applied Psychology, 82*(5), Table 1, p. 629. Copyright 1997 by the American Psychological Association.

1.3.1.2 Level 1b: Competency Model

Although job analysis has historically served as the cornerstone of any selection test development process, more recently, competency modeling has started to come into practice. A *competency model* (CM) is a collection of the KSAs, which are relevant for successful performance on a job. In a competency model, these individual KSAs are what define the range of job-related competencies (Campion, Fink, Ruggeberg, Carr, Phillips, & Odman, 2011). Table 5 summarizes several of the important differences that exist between job analysis and competency modeling.

The movement toward competency models is based, in part, on the observation that specific behavioral requirements for any given job may change over time. Thus, to avoid constantly revising job analyses and job descriptions, it may be more useful to focus on a slightly higher level of abstraction (i.e., competencies). The idea is that individuals with the requisite competencies are also likely to meet the more specific behavioral requirements that might be derived from any given job analysis.

In addition, where a job analysis is likely to yield a list of KSAOs hypothesized to underlie successful levels of job performance, a CM is typically designed to distinguish "star performers" from "average performers." Thus, in critical missions, it may be more important to identify individuals likely to 'greatly exceed' expectations vs. 'likely to exceed' minimum expectation thresholds. Another important distinction is the process of developing a CM. Instead of starting on the front-line and asking those previously in missions what tasks they performed, those at the top, perhaps officers, might first develop the competencies deemed important (i.e. top-down versus bottom-up approach to decomposing the job). Table 6 provides a list of best practices in competency modeling (Campion et al., 2011). After the job analysis or competency model is complete, this information should be used to inform the constructs of interest to be measured.

Table 5. Description of Competency Models and Key Differences between Competency Models and Job Analysis

1. Executives typically pay more attention to competency modeling.
2. Competency models often attempt to distinguish top performers from average performers.
3. Competency models frequently include descriptions of how the competencies change or progress with employee level.
4. Competency models are usually directly linked to business objectives and strategies.
5. Competency models are typically developed top down (start with executives) rather than bottom up (start with line employees).
6. Competency models may consider future job requirements either directly or indirectly.
7. Competency models may be presented in a manner that facilitates ease of use (e.g., organization-specific language, pictures, or schematics that facilitate memorableness).
8. Usually, a finite number of competencies are identified and applied across multiple functions or job families.
9. Competency models are frequently used actively to align the HR systems.
10. Competency models are often an organizational development intervention that seeks broad organizational change as opposed to a simple data collection effort.

Note. Reproduced from "Doing competencies well: Best practices in competency modeling," Campion, Fink, Ruggeberg, Carr, Phillips, & Odman, 2011, *Personnel Psychology, 64*(1), Table 1, p. 227.

Table 6. Best Practices in Competency Modeling

Analyzing Competency Information (Identifying Competencies)
1. Considering organizational context
2. Linking competency models to organizational goals and objectives
3. Start at the top
4. Using rigorous job analysis methods to develop competencies
5. Considering future-oriented job requirements
6. Using additional unique methods
Organizing and Presenting Competency Information
7. Defining the anatomy of a competency (the language of competencies)
8. Defining levels of proficiency on competencies
9. Using organizational language
10. Including both fundamental (cross-job) and technical (job-specific)
11. Using competency libraries
12. Achieving the proper level of granularity (number of competencies and amount of detail)
13. Using diagrams, pictures, and heuristics to communicate competency models to employees
Using Competency Information
14. Using organizational development techniques to ensure competency modeling acceptance and use
15. Using competencies to develop HRs systems (hiring, appraisal, promotion, compensation)
16. Using competencies to align the HR systems
17. Using competencies to develop a practical "theory" of effective job performance tailored to the organization
18. Using information technology to enhance the usability of competency models
19. Maintaining the currency of competencies over time
20. Using competency modeling for legal defensibility (e.g., test validation)

Note. Reproduced from "Doing competencies well: Best practices in competency modeling," Campion, Fink, Ruggeberg, Carr, Phillips, & Odman, 2011, *Personnel Psychology, 64*(1), Table 2, p. 230.

1.3.1.3 Level 1c: Construct Specification/Definition

As noted above, test validation begins with the purpose of testing and the proposed interpretation of test scores. Any such interpretation of test scores necessitates "specifying the construct the test is intended to measure" (p. 11; AERA et al., 2014). Thus, test development and validation efforts should be built upon a strong foundation, one anchored to a good *construct definition.* Strong construct definitions are essential to validation efforts because they are important to developing our operationalizations (i.e. our measurement) of our focal constructs (Podsakoff, MacKenzie, & Podsakoff, 2016).

As previously discussed, one crucial step in the validation of a test or assessment is ***divergent (or discriminant) validity*** with tests designed to measure dissimilar constructs. However, when construct definitions are not clear, and thus the construct one intends to measure may actually be more similar than intended to other constructs, it is more difficult to gather this evidence (Podsakoff et al., 2016). Thus, a good construct definition explains what is unique about this construct (i.e., what the construct "is"), but also what differentiates it from similar constructs (i.e., what the construct "is not"). Similarly, we know that gathering evidence for ***criterion***

validity is a key step in this process. However, if the construct was not well defined when developing the test, the test may not adequately sample from the relevant construct domain (i.e., measurement deficiency) or it may inappropriately sample from irrelevant domains (i.e., measurement contamination). Any construct irrelevant variance in our measures may attenuate (i.e., when the irrelevant variance is error) or systematically bias (i.e., when the irrelevant is systematic) relationships with external variables (Podsakoff et al., 2016).

Podsakoff and colleagues (2016) offered practical suggestions for developing good conceptual definitions (see Table 7). These suggestions span for four steps (albeit sometimes iterative and overlapping):

1. *The first step is to "identify potential attributes by collecting a representative set of definitions" (Podsakoff et al., 2016, p. 169).*

As an example, let's say that one is trying to develop a measure of "grit" because it might be a good predictor of performance outcomes in certain missions. In specifying the construct "grit", the first step might be to survey relevant literature for existing definitions of grit (and related constructs). One may also want to interview relevant officers or personnel who have had experience on similar missions and ask what "having grit" means to them or ask them to identify instances where they saw grit in action.

2. *The second step is to "organize the potential attributes by theme and identify any necessary and sufficient or shared ones" (Podsakoff et al., 2016, p. 169).*

In order to be able to identify similarities and differences across definitions, one must do a sufficiently thorough initial search. Continuing with the previous example, one might notice themes around being careful, or being organized, or being thorough. Thus, one may determine that conscientiousness is a key attribute of grit (Duckworth, Peterson, Matthews, & Kelly, 2007). After attributes like these have been identified; it is useful to examine the list of attributes to determine which are necessary and which are sufficient for "grit". Did the attribute "conscientiousness" come up in every definition? What attributes are sufficient when combined with one another? Once this step is complete, one might look for a set of attributes that are most important for the construct of interest and ensure that these are used in the definition. Note that the term theme is used to emphasize the inductive nature of the exercise (i.e., reading and reflecting on extant definitions of grit to identify themes that emerge across definitions). When multiple definitions seem to triangulate on a common theme, then it probably makes sense to consider that theme as a possible defining attribute of the construct. For example, definitions of grit might include descriptions of behavior that reflect tendencies to be dependable, organized, or persistent. A psychologist might look across various conceptualizations of "grit" that include these types of behavior and identify a "theme" around conscientiousness. This process would likely reveal additional themes linked to the grit construct. The psychologist could then formally consider the extent to which these themes should be considered key attributes of the hypothesized grit construct. For example, attributes may be evaluated to determine which ones are sufficient for adequately representing the grit construct.

3. *The third step is to "develop a preliminary definition of the concept (Podsakoff et al., 2016, p. 169).*

Once the key attributes of the construct have been determined, the third step is to develop a preliminary definition. This stage is made up of various substages. The first step Podsakoff and colleagues (2016) recommended was to specify the "type of property the concept represents and the entity to which that property applies" (p. 184). In regards to our example, grit would be an intrinsic characteristic (the property) of a person (the entity). The authors give various examples of properties including intrinsic characteristics, thoughts, feelings, perceptions, actions or performance metrics. Other entities might include tasks, processes, relationships, teams, organizations etc. One important note, if we are continuing with the grit example, is that each subdimension, should be clearly defined. So, whereas you might define grit as being made up conscientiousness and other facets, conscientiousness needs also be defined specifically as related to grit. Another step in this stage of the process is to make explicit if the construct is stable (both over time and across situations; Podsakoff et al., 2016). Because grit, is considered an intrinsic characteristic, like other personality traits, we would assume it should remain stable over time and across situations. The next substage is to differentiate the construct of interest from related constructs. Using grit, for example, it would be important to establish how grit is different from achievement motivation. Finally, one should start to consider the construct's antecedents and consequences.

4. *The final step is to "refine the conceptual definition of the concept" (Podsakoff et al., 2016, p. 169).*

Once a preliminary definition has been constructed, it must be refined. The main way that Podsakoff and colleagues (2016) suggest doing this is to rid the definition of any ambiguity. Specifically, "continue to ask the question about more specific aspects of the definition until no more ambiguity exists" (p. 187). During this stage, one may also want to consult subject matter experts for opinions on the current form of the definition. Again, this is an iterative process, so one may need to work back through these stages even after this stage is complete.

Table 7. Summary of Stages for Developing Good Conceptual Definitions

Stage 1. Identify potential attributes by collecting a representative set of definitions	Stage 2. Organize the potential attributes by theme and identify any necessary and sufficient or shared ones	Stage 3. Develop a preliminary definition of the concept	• Stage 4. Refine the conceptual definition of the concept
• Search the dictionary • Survey the literature • Interview experts, colleagues and/or practitioners • Conduct focus groups • Use direct (structured) observation • Use case studies • Compare the concept with its opposite pole • Examine current operationalizations of the concept or think about how the concept might be operationalized	• Condense the attributes in Step 1 into a reduced set • Identify any attributes that are necessary and sufficient to the definition of the concept or (alternatively) identify shared attributes across subsets of cases • Try to identify: (a) a theoretical framework that helps organize the attributes along their defining dimensions; and/or (b) the criteria that should be used to decide which attributes to include (and which to exclude) in the concept's definition	• Describe the type of property the concept represents and the entity to which it applies • Describe the necessary and sufficient attributes of the concept • Specify the dimensionality of the concept • Specify the stability of the concept • Specify how the attributes of the focal concept differ from the attributes of other, related concepts • If possible, identify some of the antecedents and consequences of the concept	• Ask "What do we mean by that?" until all of the ambiguity in the words used to define the focal concept have been resolved • Reduce jargon by: (a) playing the role of a journalist who is asked to write a description of the focal concept; and/or (b) imagining trying to explain the concept to someone learning English • Solicit feedback from peers

Note. Adapted from "Recommendations for creating better concept definitions in the organizational, behavioral, and social sciences," Podsakoff, MacKenzie & Podsakoff, 2016, *Organizational Research Methods, 19*(2), Figure 1, p.182.

1.3.1.4 Level 1d: Test Development

After the focal construct has been identified and defined, researchers can begin writing items (or building tasks) that they believe adequately sample the relevant construct domain. The specific form that an item takes will largely depend on the nature of the latent construct. Items designed to measure specific knowledge, skills, or abilities may written in a manner that lends itself to dichotomous scoring (e.g., correct-incorrect). In contrast, items designed to measure needs, motives, values, interests, attitudes, personality traits, and/or "other" characteristics could be

written in a manner that lends itself to either dichotomous scoring (e.g., yes-no; accurate-inaccurate; true-false) or polytomous scoring (e.g., 5 or 7 point Likert-type scales). A detailed summary of best practices for all item types is beyond the scope of the current report. However, the following resources may prove as useful starting points: Hinkin (1998), Lievens and Sackett, 2007; Lozano, Garcia-Cueto, & Muniz (2008), McDonald (2000; especially chapter 2), Schwarz (1999), and Wakita, Ueshima, and Noguchi (2012). As these, and other sources document, the nature of the underlying construct, the purpose of the test, and the format of the item will directly impact any recommendations for writing "optimal" items. Nevertheless, irrespective of the specific construct or item format that is being used, it is possible to make a few general recommendations concerning item construction:

- *Researchers should be sure to evaluate the reading-level of their items to confirm that they are appropriate for members of the target population.*

- *Researchers should be sure to avoid using language that may viewed as inappropriate or offensive by prospective test takers.*

- *Researchers using self-report surveys should take care to avoid "double-barreled" items (e.g., I don't trust my squad members or my platoon leader).*

- *Researchers building tests that reliably measure the entire range of the construct continuum, θ, should include items with a range of difficulty levels.*

    - *It is possible to adjust item-level difficulty by adjusting the attractiveness of the "incorrect" (or distractor) solutions when items are used to measure cognitive traits.*

- *Researchers building tests to discriminate at a specific level of the construct continuum (i.e., a specific cut-score) should include items with difficulty levels in the vicinity of the desired cut-score.*

- *Researchers develop a sufficiently large pool of initial items.*

    - *It is not uncommon for 1/2 to 1/3 of the items comprising an initial item pool to be problematic and require revision or removal based on initial item analyses; especially, if this is the first-time items are being written to measure the focal construct.*

1.3.1.5 Level 1e: Pilot Testing

Once an initial set of items has been developed, researches should undertake a preliminary pilot testing of the items (Hinkin, 1998). The purpose of this pilot study is a) to verify that the items have sufficient variability to warrant inclusion in subsequent studies, b) provide an opportunity to collect some initial content validity evidence, and c) to confirm that the test is clear and accessible to members of the population(s) of interest.

- *After an initial draft of the test has been built, researchers should submit the test items to a content validity analysis (see Validity Evidence Based on Test Content).*

- *Researchers should also verify that the items are written at a level that will be clear and understandable to members of the target population.*

- *Researchers should also verify that items are free from biased or offensive language or themes.*

- *Researchers may wish to interview or debrief the pilot study sample to obtain information about problematic, confusing, or offensive items.*

- *Researchers interested in obtaining validity evidence based on response processes may also ask members of the pilot study sample to engage in a verbal protocol analysis as they complete the test. This verbal protocol analysis represents a "think out loud" exercise that enables researchers to better understand how respondents approached and completed the test.*

- *Finally, researchers should revise or remove items that were flagged as problematic (little if any variability in response patterns; confusing to test takers; etc.).*

1.3.2. Level 2 Validation-Concept Exploration

"At this stage the test or measure is directly utilized on the problem source (i.e. - specific career field attrition problem, AFSC specific training gap) in order to see if it has any positive effect at reducing the problem or meeting the need. This level is about "proof of concept" on pre-existing samples/sources within the affected pipelines-typically training programs." (p. 25; AFEAO, 2010).

This stage of validation encompasses several sub-stages including data collection, formal item/test analysis, and accumulation of validity evidence based on relationships with other variables. Data should be collected on a sufficiently large sample that is representative of the focal population (Hinkin, 1998).

1.3.2.1 Level 2a: Sample Size Determination

At this stage in the process, data should be collected on a large enough sample to permit the accurate estimation of statistics to evaluate the items/test. Specifically, researchers will be conducting a formal item analysis based on classical test theory, item response theory, or some combination thereof. The specific sample sizes needed to obtain stable point estimates will vary as a function of the different statistics being estimated. However, irrespective of whether CTT, IRT, or some hybrid are used, researchers will likely compute a combination of estimates: a) item and test means, b) item and test variances, c) inter-item covariances, d) item difficulty, e) item discrimination, f) item-total correlations, g) item-criterion correlations, h) factor loadings and factor intercorrelations, i) test-criterion correlations, j) convergent/divergent correlations between test scores and external variables, k) item characteristic curves, and l) item information curves, and/or m) estimates of reliability.

A number of heuristics have been developed concerning sample sizes needed to conduct a particular analysis. For example, Hinkin (1998) reviewed heuristics related to sample sizes

needed for conducting factor analysis, while de Ayala (2009) included suggestions for IRT item calibration sample sizes for each of the different IRT models that were reviewed. Alternatively, researchers might ask how large of a sample is needed to provide precise and stable estimates of the relevant statistics (see Tonidandel, Williams, & LeBreton, 2015 for a review for correlation, regression, and factor analysis). More recently, Mair (2018) suggested that the ideal approach to estimating minimum sample sizes for IRT analyses involves the use of Monte Carlo simulations and he provided an example in R.

- *Researchers should consult with appropriate sources (see previous paragraph for references) when determining the minimum sample sizes necessary for conducting item and test evaluations.*

1.3.2.2 Level 2b: Evaluation of Items and Development of the Initial Test

Once data has been collected from a large and representative sample, researchers should conduct ***item analyses*** based on Classical Test Theory (CTT) or Item Response Theory (IRT) (see subsequent sections of this report); or, researchers may opt to undertake an item/test analysis that is based on an integration or combination of CTT and IRT (see, for example, Smith, Hoffman, & LeBreton, 2020).

- *Researchers should conduct an item analysis to identify both problematic and non-problematic items. The latter set of items will be used to form an initial (developmental) version of the test.*

After using item analysis to develop the initial (developmental) version of test, researchers should then accumulate additional validity evidence based upon how test scores correlate with other variables (see previous section of the report concerning sources of validity evidence). At a minimum:

- *Researchers should accumulate initial evidence for inferences linking test scores with measures of both related and unrelated constructs (i.e., evidence of convergent and divergent validity).*

- *Researchers should accumulate initial evidence for inferences linking test scores with organizationally valued outcomes/criteria (e.g., performance, attrition, attitudes, other job-relevant behaviors).*

1.3.3. Level 3 Validation-Program Definition and Risk Reduction

"Level 3 involves exploring test utility in a broader context and, at the same time, ensuring that the predictive validity of this proposed test is not already covered by other existing or operational tests" (p. 25; AFEAO, 2010).

1.3.3.1 Incremental Importance/Validity

During this stage of validation, the focus is on continued efforts to accumulate validity evidence by collecting data from broader contexts/samples. A particular emphasis is placed on establishing the incremental predictive validity of the test by demonstrating it explains unique

variance in an outcome or criterion measure. As noted above, ***incremental validity*** is typically tested using hierarchical regression analysis where existing measures are included in Step 1 and the new measure is added in Step 2. The significance test for incremental validity is obtained either by examining p-value for the t-test associated with the unstandardized regression coefficient for the new measure or by examining the p-value for the F test associated with the change in R^2 estimated by subtracting the R^2 from Step1 from the R^2 obtained in Step 1– these two significance tests (and the accompanying p-values) are equivalent. Establishing incremental validity is useful because it allows researchers to verify that the new measure is not statistically redundant with the existing set of measures. As LeBreton, Hargis, Griepentrog, Oswald, and Ployart (2007) noted:

> "…I-O psychologists often statistically evaluate new variables by examining the importance of those new variables compared to an existing set of variables. One definition of variable importance emphasizes the incremental validity of the new measure, which we call incremental importance. This definition of importance was suggested by Darlington (1968) with his usefulness statistic. Incremental importance is valuable because it ensures that the variable of interest is tapping unique variance in the criterion variable above and beyond that of the other variables in the regression model (Cronbach & Gleser, 1957; Sechrest, 1963)" (p.476).

Thus, by estimating incremental importance, researchers are able to confirm that a new measure is not statistically redundant with an old measure. However, any variability that the new measure shares with the criterion and the existing test battery is credited to the tests in that battery (LeBreton et al., 2007). Typically, researchers do not strive to build new tests that will be highly correlated with existing elements of a test battery. However, a new variable may nevertheless be partially correlated with elements of a test battery due to measurement similarities or nomological proximity between the new and existing tests. LeBreton et al. summarized:

> "…any criterion variance predicted by both the new variable and the existing set of variables is automatically "credited" toward the latter. Thus, an incremental validity analysis might lead one to make incorrect or misinformed decisions about the relative efficacy of the new variable. As such, it is possible that a new measure [of a new construct] might yield relatively small increments in prediction (e.g., $\Delta R^2 = .02$) but that the overall contribution that this new [measure] makes to the R^2 is as high as (or higher than) the other predictors in the model" (p. 477).

1.3.3.2 Relative importance/validity

To address this concern, LeBreton et al. (2007) recommended that researchers supplement any tests of incremental importance/validity with additional tests of the new measure's ***relative importance,*** which they defined as "the contribution each predictor makes to the R^2, considering both its unique contribution and its contribution in the presence of other predictors" (p. 477). There are multiple indices of relative importance, but the most commonly used measures are dominance analysis and relative weight analysis (see Johnson & LeBreton, 2004; Krasikova, LeBreton, & Tonidandel, 2011; Tonidandel & LeBreton, 2011) for reviews of relative importance statistics.

By reanalyzing data from several published articles, LeBreton et al. (2007) demonstrated how small changes in incremental validity could mask more meaningful contributions to the overall prediction of the criterion. For example, they demonstrated that new biodata measures "accounted for small-to-moderate increases in the model R^2. However, the relative importance analyses revealed that not only did these scales add increments to the regression, *they repeatedly emerged as the most important predictors of performance"* (p. 488). Over the last 15 years, substantial progress has been made in the development and refinement of tests of relative importance (e.g., extension to multivariate criterion spaces; dichotomous criterion variables; significance testing; etc.). The interested reader is directed to: LeBreton, Ployhart, and Ladd (2004); LeBreton and Tonidandel (2008); LeBreton, Tonidandel and Krasikova (2013); Tonidandel and LeBreton (2010, 2011, 2015); and Tonidandel, LeBreton, and Johnson (2009).

- *Researchers should establish the incremental validity/importance of new measures using traditional hierarchical regression analyses.*

- *Researchers are encouraged to supplement tests of incremental importance with tests of relative importance-namely relative weight analysis or dominance analysis.*

    - *The combination of such tests is likely to provide a more thorough and complete understanding of the value that a new measure has when predicting relevant criteria.*

1.3.4. Level 4 Validation-Engineering and Manufacturing Development

"This level involves giving the test or measure at the broadest level of testing which is the general applicant population. Showing that the test or measure can improve selection or classification (reduce or solve the identified problem or need) on the applicant population means time and money can be saved by implementing the test at the earliest stage of the personnel life scores are developed through data collection and analysis" (p. 26; AFEAO, 2010). *cycle. At this level, normative data is collected and potential cutoff/qualification*

This stage of validation involves the continual accumulation of validity evidence by using the test to predict relevant organizational criteria. At this stage, researchers collect data from sufficiently broad and representative samples, so as to allow the creation of potential cutoff scores or minimum qualification scores on the test. However, it is important that any cutoff scores are coherently developed and clearly documented. It is also important that researchers be mindful of how cutoff scores may adversely impact members of protected classes.

- *Researchers may develop cutoff values for test scores, but care should be taken to avoid values that will likely engender adverse impact.*

1.3.5. Level 5 Validation-Production, Deployment, Operational Support, and Monitoring

"This level means the test or measure is now in operational use and personnel decisions can be made based upon the test or measure. Initial Operational Test and Evaluation (IOT&E) with the initial norms developed in Level 4 are validated and ongoing monitoring and evaluation occurs throughout the operational life cycle of the test or measure" (pp. 26-27; AFEAO, 2010).

In this final stage of development, researchers will actively monitor the use of the test, and the performance of any established test norms and cutoffs. Researchers are also encouraged to continue accumulating validity by examining correlations between the test scores and other, external variables (i.e., criterion validity; convergent/divergent validity).

- *After a test has been developed, subjected to substantial validation efforts, and is in operational use, researchers should continue to monitor the performance of the test (including any test norms or cutoffs) and continue to accumulate validity evidence.*

    - *The continued monitoring and validation of the test ensures that there is no "drift" over time in the validity of the inferences being drawn from test scores.*

1.4 Using Classical Test Theory to Evaluate Items and Build Tests

1.4.1. Symbols and Notation

The following symbols and notation, based largely on Gulliksen (1950); will be used throughout the remainder of this report.

- X, Y, Z = observed/raw scores
- x, y, z = observed/raw scores in deviation score format; observed/raw scores minus the mean
- i and j = subscripts denoting different examinees
- g and h = subscripts denoting different items or tests
- N = total number of examinees
- n = number of examinees in a subgroup
- K = total number of items or number of tests in a test battery
- k = number of items in subtest
- T = unobserved/latent true score
- t = unobserved/latent true score in deviation score format; true score minus the mean
- E = unobserved/latent score corresponding to random measurement error
- e = unobserved/latent error score in deviation score format
- $M, \bar{X}, \bar{Y}$ = mean or expected value
- S, sd = sample standard deviation
- r = sample correlation coefficient
- μ = population mean
- σ = population standard deviation
- ρ = population correlation coefficient
- θ = latent construct being measured by a particular set of items or tests

1.4.2. Overview of Classical Test Theory

1.4.2.1 Primary Assumptions of Classical Test Theory

Classical test theory (CTT) rests upon a set of important assumptions concerning the patterns of relationships that are presumed to exist (or not) between observed scores, true scores, and error scores. The first assumption is given by:

$$X_i = T_i + E_i \quad (5),$$

where, X_i refers to the observed score of person *i* on test (or item) *X*, T_i refers to the unobserved true score for person *i* on this test, and E_i refers to the unobserved error score for person *i* on this test. Thus, the first assumption of classical test theory is that any given person's observed score may be represented as a unit-weighted linear composite of his or her true score and error score. This equation introduces the impossible task of trying to solve for two unknown values (T_i and E_i) using only a single equation with one known value (X_i; Gulliksen, 1950). However, by collecting additional data from other individuals and imposing additional assumptions, we are able to develop items using CTT. The additional assumptions include:

$$M(X) = T \quad (6),$$

$$\rho_{ET} = 0 \quad (7),$$

$$\rho_{E_1E_2} = 0 \quad (8), \text{ and}$$

$$\rho_{E_1E_2} = 0 \quad (9).$$

Briefly, equation 6 states that observed scores have a mean or expected value equal to the true score. Equation 7 states that error scores are uncorrelated with true scores. Equation 8 states that random errors on two parallel tests will be unrelated to one another; and finally, equation 9 states that the errors on one test will be uncorrelated with the true scores on another test (see Allen & Yen, 1979; Lord & Novick, 1968).

1.4.2.2 Psychometric Item Types

Two items (or tests) are defined as ***parallel measurements*** of a construct when they share a common true score (T_1=T_2), their errors are linearly independent of one another ($\rho_{E_1E_2} = 0$), and the error variances are equivalent ($\sigma^2_{E1} = \sigma^2_{E2}$; Lord & Novick, 1968). Relaxing the assumption that these two items must have equivalent error variances yields *tau-equivalent measurements* (i.e., true score equivalent measurements). Finally, relaxing both the assumptions that the error variances are equivalent and the true scores are equivalent yields *congeneric measurements.* Ideally, one has parallel (or at least tau-equivalent items). When we meet this assumption, we are able to compute traditional estimates of internal consistency reliability (e.g., Coefficient Alpha). However, when we have congeneric items, alpha will underestimate reliability. With congeneric items, one is advised to compute a reliability estimate that does not make the strict assumption of true-score equivalence (e.g., MacDonald's Coefficient Omega). Most of classical test theory assumes access to parallel items or tau equivalent items.

1.4.2.3 Conclusions Drawn from Classical Test Theory

Using equations 5-9, it is possible to derive the following:

$$M(E) = 0 \quad (10),$$

which means that the mean or expected value of random errors is zero (i.e., errors are just as likely to be positive as negative, and in the expectation, will equal zero).

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2 \quad (11).$$

If *g* and *h* are parallel measures, then

$$M(T_g) = M(T_h) \quad (12),$$

$$S(T_g) = S(T_h) \quad (13), \text{ and}$$

$$r_{T_g,T_h} = 1.0 \quad (14).$$

Equation 11 states that the total observed variance of *X* is equal to a unit-weighted composite of the true score variance and the error score variance. Equations 12 and 13 state that true scores on parallel tests have equal means and equal variances. Finally, equation 14 states that true scores on parallel tests will be perfectly correlated with one another.

1.4.2.4 Reliability Index, Reliability Coefficient, and the Standard Error of Measurement

Reliability refers to the consistency or stability of measurements. Under CTT, information about the reliability of items and tests is provided by the reliability index, reliability coefficient, and the standard error of measurement.

Reliability index. The correlation between observed scores and true scores is referred to as the reliability index. Using equations 5 through 14, it is possible to show:

$$\rho_{XT} = \frac{\sigma_T}{\sigma_X} \quad (15).$$

Thus, the correlation between true scores (T) and observed scores (X) is equal to the ratio of the true score standard deviation to the observed score standard deviation. Although the reliability index is psychometrically interesting, it is practically useless because true scores are unobserved and thus cannot be used to estimate this correlation coefficient (Crocker & Algina, 1986).

Reliability coefficient. A more practical correlation may be obtained using scores from two parallel measures, g and h. This correlation is referred to as a reliability coefficient and is equal to:

$$\rho_{X_g X_h} = \frac{\sigma_T^2}{\sigma_X^2} \quad (16),$$

indicating that that reliability is equal to the ratio of true variance to total variance. Recall that the total variance is equal to the sum of the error variance and true variance. Thus, equation 16 may be rewritten as:

$$\rho_{X_g X_h} = \frac{\sigma_T^2}{\sigma_T^2+\sigma_E^2} \quad (17).$$

The reliability coefficient may simply be denoted ρ_{XX}. When ρ_{XX}= 1, the test is perfectly reliable indicating that true observed scores provide perfect approximations of true scores (i.e., 100% of the observed score variance is attributed to variance in true scores). When ρ_{XX}= 0, the test is perfectly unreliable, suggesting that 100% of the observed variance is attributed to random error variance. Finally, we see that the reliability coefficient is equal to the square of the reliability index:

$$(\rho_{XT})^2 = \rho_{XX} \quad (18).$$

Standard error of measurement. Although we typically think of reliability coefficient as the ratio of true variance to total variance, we can use equations 5 through 14 to show that the reliability coefficient is also equivalent to:

$$\rho_{XX} = 1 - \frac{\sigma_E^2}{\sigma_X^2} \quad (19).$$

Solving equation 19 for σ_E^2 yields:

$$\sigma_E^2 = \sigma_X^2(1 - \rho_{XX}) \quad (20).$$

Taking the square root of equation 20 yields the *standard error of measurement* (i.e., standard deviation of the error scores):

$$\sigma_E = \sigma_X\sqrt{1 - \rho_{XX}} \quad (21).$$

Under the CTT model, the standard error of measurement is a single (i.e., constant) value is uniformly applied to all examinees and may be used to build confidence intervals around examinees' observed scores. Such confidence intervals may be useful for estimating "how far the true score may lie from an observed score for an average examinee in the population" (Crocker & Algina, 1986, p.124).

1.4.3. Psychometric Evaluation of Items and Tests

1.4.3.1 Item-Level Evaluations

Item difficulty. In classical test theory (CTT), ***item difficulty*** is defined as the relative frequency of individuals endorsing the "correct" or "keyed" item response alternative (Allen & Yen, 1979; Ghiselli, Campbell, & Zedeck,1981; Lord & Novick, 1968). The label, item difficulty, is a bit of a misnomer as higher item difficulty values actually imply an easier item.

For example, an item difficulty of .10 indicates that only 10% of the sample endorsed the correct response, thus indicating a relatively difficult item. In contrast, an item difficulty of .90 indicates that 90% of the sample endorsed the correct response, thus indicating a relatively easy item. As Allen and Yen (1979) noted, "The words ***difficulty*** and ***correct*** are best suited for discussions of ability or achievement tests. If a personality test is being developed, a "correct" response is a response that counts toward the trait and the "difficulty" of an item reflects the popularity of the "correct" response-that is, the proportion of examinees who chose this response" (pp. 120-121). With dichotomously scored items, the difficulty values are simply the item means, labeled as *p*-values, to denote the probability of endorsing the correct or keyed item response. In addition, the phrase "item split" is used to denote the likelihood of endorsing the correct response relative to the incorrect response. For example, an item with a .05-.95 split, indicates 5% of the sample endorsed the "correct" answer and 95% selected the incorrect answer.

Item discrimination. In CTT, item discrimination is defined as the extent to which an item is effective at distinguishing between different levels of the focal construct (e.g., relatively high levels of cognitive ability vs. relatively low levels of cognitive ability). There are several ways to estimate item discrimination. The ***item-discrimination index*** for a given item, *g,* is given by the differences in *p*-values between individuals with high scores (i.e., upper end of the distribution) and individuals with low scores (i.e., lower end of the distribution):

$$dg = p_u - p_l \quad (22),$$

where p_u is the proportion of individuals in the upper group who correctly answered item *g,* and p_l is the proportion of individuals in the lower group who incorrectly answered item *g.* In order to compute d_g, researchers must identify cut-points that will be used to create the groups comprising the upper and lower score groups. As Allen and Yen (1979) noted, "Upper and lower ranges generally are defined as the upper and lower 10% to 33% of the sample, with examinees ordered on the basis of their total test scores." If test scores are normally distributed, it is recommended that researchers set cut-points to include the upper 27% and lower 27% of examinees.

An alternative index of discrimination is provided by the correlation between an item and the total test score-***the item-total correlation.*** A positive correlation indicates a properly functioning item (i.e., as scores on the test increase, so too does the probability of endorsing the correct item). For dichotomously scored items, there are two options for computing item-total correlations: point-biserial correlations and biserial correlations. The ***item-total point-biserial correlation*** is simply a product-moment correlation and represents the appropriate statistic to compute when one has a continuous variable (e.g., total test score) and a truly dichotomous variable (e.g., sex). One limitation of this statistic is that it does not assume a typical range of values, but rather is constrained to take on values between -.80 and +.80. The maximum correlation of ~.80 is only observed when the dichotomously scored variable has a p-value (i.e., item difficulty) near .50 (Nunnally & Bernstein, 1994). As the p-values (i.e., item difficulties) become more extreme, the functional range of item-total point-biserial correlations becomes further restricted. For example, items with a p-value of .10, have a maximum point-biserial correlation of ~.50.

In contrast, if one is able to assume that the underlying construct distribution is normally distributed, it is appropriate to estimate the ***item-total biserial correlation,*** which assumes the typical range of values for correlation coefficients (i.e., -1.0 to +1.0). Some researchers believe that the added benefits of the biserial correlation are offset by it having greater sampling error, which is less of a concern with access to large samples. Lord and Novick (1968) summarized the coefficients thusly, "The point biserial correlation gives the actual product moment correlation between test score, or external criterion, and item. We may view the biserial simply as another measure of association, one different from the product moment correlation. The biserial is widely used because it is hoped that the biserial will demonstrate a type of invariance from one group of examinees to another not provided by the point biserial" (p. 341). Ultimately, it comes down to the comfort level of the researcher. If there is a continuous, but skewed, latent distribution then using the point-biserial will potentially yield a biased estimate because it assumes the latent distribution is not continuous. However, so too would using the biserial correlation, because it assumes the latent distribution is normal. In summary, if a researcher has a dichotomously scored item (e.g., correct vs. incorrect) and the underlying latent construct is presumed to be continuously and normally distributed, the biserial correlation is preferred. In contrast, if a researcher has a dichotomously scored item (e.g., White vs. Black) and the latent construct is presumed to be dichotomous, then the point-biserial correlation is preferred. When in doubt, one could always compute and report both types of correlations.

Another concern with item-total correlations is that estimates will be larger whenever the focal item is included as part of the overall composite (i.e., total) score. Thus, some researchers prefer to compute the "corrected" item-total correlation between an item, g, and a total test score computed after first excluding the focal item from the total test score.

Finally, with multiple choice tests, it is useful to examine how examinees in different ability groups endorse each of the item responses. For cognitive tests, the incorrect answers are referred to as distractors. By endorsement patterns for both the keyed responses and the distractor responses, it is possible to identify response options that may require revisions. For example, if a distractor is regularly endorsed by members of the upper ability group but rarely endorsed by members of the lower ability group, then it is in need of revision.

Item reliability index. In CTT, the *item reliability index* is defined as the product of the correlation between the focal item and the total score (i.e., item-total correlation) and the standard deviation of the focal item. The item reliability index essentially weights item discrimination by the magnitude of item variability.

- *When the goal is to select items for a test that will maximize estimates of internal consistency reliability, then researchers should select items with a range of standard deviations and positive item-total correlations (see pp. 125-126; Allen & Yen, 1979).*

Item validity and item validity index. In CTT, ***item validity*** is defined as the extent to which item responses predict some relevant criterion variable. The simplest estimates of item validity are simply bivariate correlations between each item and the criterion variable, denoted ***item-criterion correlations.*** Different types of coefficients are appropriate when estimating item-criterion correlations, conditional on the type of item and type of criterion variable involved (e.g., biserial, point-biserial, polyserial, polychoric, tetrachoric, phi). In addition, researchers may

also compute the ***item validity index*** which is defined as the product of the correlation between the focal item and the criterion variable (i.e., item-criterion correlation) and the standard deviation of the focal item. The item validity index essentially weights item validities by the magnitude of item variability.

- *When the goal is to select items that will maximize the predictive validity the test, then researchers should select items with a wide range of item-reliability index values and positive values for item-criterion correlations (see pp. 125-126; Allen and Yen, 1979).*

1.4.3.2 Test-Level Evaluations

In addition to undertaking item analyses, researchers will want to examine the psychometric quality of the items as a set, that is researchers should also plan to undertake psychometric and validity analyses as part of the test level evaluation process. Under the CTT framework, test-level evaluations may include the estimation of reliability coefficients, fitting data to exploratory and/or confirmatory factor analyses, and the accumulating validity evidence for the test by examining patterns of correlations with external variables.

Internal consistency. In evaluating the ***internal consistency reliability*** (i.e., item interrelatedness; Cho & Kim, 2015) of a test, the most frequently used estimate has been *Cronbach's coefficient alpha* (Cronbach, 1951). However, a number of researchers have long lamented the misuses and misinterpretations of coefficient alpha (Cho & Kim, 2015; Cortina, 1993; McNeish, 2018). There is growing consensus that coefficient alpha should only be used when its accompanying statistical assumptions are likely to be met. As McNeish (2018) noted, there are four basic assumptions that must be met in order to justify the use of coefficient alpha as an estimate of internal consistency reliability. First, the items should be considered tau-equivalent, implying that "each item on a scale contributes equally to the total scale score" (p. 415). Analytically, the tau equivalence assumption could be tested by examining the factor loadings in an exploratory factor analysis by extracting a single factor and examining the standardized loadings to ensure they are all roughly equivalent. Alternatively, if researchers were conducting a confirmatory factor analysis, they could simply compare two nested models. Model 1 would freely estimate all factor loadings and Model 2 would constrain all loadings to be fixed to a common estimate. A chi square difference test could be used to determine whether there was a statistically significant difference in the fit between the data and Model 1 (i.e., the congeneric item model) versus Model 2 (i.e., the tau-equivalent item model). If the chi square test is statistically significant, then the researcher may infer the congeneric item model was a better fit to the data than the more restrictive tau-equivalent item model. If the chi square test is non-significant, then the researcher may infer that the tau-equivalent model is a statistically reasonable/plausible model for explaining the data covariance structure..

The second assumption underlying the use of coefficient alpha is that the items are measured on a continuous scale that is normally distributed (McNeish, 2018). Given that most items used in psychological research are, at best, measured on an interval scale (e.g., Likert-type response scales), it is possible to base estimates of coefficient alpha using "a polychoric covariance (or correlation) matrix rather than a Pearson covariance matrix" (p. 415). Use of the polychoric matrices is predicated on the assumption that the latent construct that is the target of

measurement is normally distributed. If this assumption is tenable, then use of the polychoric matrices provides a more accurate estimate of alpha when the item scaling is discrete.

The third assumption of coefficient alpha is that the errors on a test are pairwise uncorrelated (McNeish, 2018). Unfortunately, there are many instances where researchers may unknowingly violate this basic assumption. For example, correlated errors may be engendered by "…the order of items on the scale (Cronbach & Shavelson, 204; Green & Hershberger, 2000), speeded tests (Rozeboom, 1966), transient response where feelings or opinions may change over the course of the scale (Becker, 2000; Green, 2003), or unmodeled multidimensionality of a scale (Steinberg & Thissen, 1996)" (McNeish, 2018). It is possible to examine the extent to which this assumption may be violated by requested modification indices when fitting data to a single factor confirmatory factor analysis. Researchers can examine these indices to determine whether violations of this assumption seem likely.

The fourth and final assumption underlying the use of coefficient alpha is that that the items are measuring a single, unidimensional construct (McNeish, 2018). As Cortina (1993) convincingly demonstrated, it is possible to obtain high estimates of coefficient alpha, even when the underlying data are multidimensional. Thus, the assumption of undimensionality must be met prior to using coefficient alpha to estimate internal consistency reliability. This assumption may be tested using factor analysis.

- *Researchers should only estimate internal consistency reliability using coefficient alpha after first confirming that the data appear to meet the requisite assumptions.*

When the data do not support the use of coefficient alpha, researchers are encouraged to estimate internal consistency reliability using more appropriate statistics. The most commonly recommended alternative is composite reliability based upon *coefficient omega* (McDonald, 2000). McNeish (2018) reviewed several different variants of omega that are appropriate when items are congenric (i.e., ***omega total)*** and when the items are not truly unidimensional but rather may be measuring "additional minor dimensions" (i.e., *omega hierarchical*). For additional details on when alpha or alternatives may be most appropriate, the reader is directed to Cho (2016), Cho and Kim (2015), Cortina (1993), McDonald (2000), and McNeish (2018). For a discussion of how to select appropriate estimates of reliability as a function of item type (i.e., parallel vs. tau-equivalent vs. congeneric) and scale dimensionality (i.e., unidimensional vs. multidimensional), the reader is strongly encouraged to consult Cho (2016); see also alternative statistics discussed in McNeish (2018).

- *When the assumptions of coefficient alpha are not tenable, internal consistency reliability should be estimated using appropriate alternatives.*
    - *When data are multidimensional, researchers should consider estimating the multidimensional version of omega (Cho & Kim 2015; McDonald, 2000; McNeish, 2018) or stratified alpha (Cho & Kim, 2015).*
    - *When data are congeneric, researchers should consider estimating the unidimensional version of omega (Cho & Kim, 2015) or coefficient H (McNeish, 2018).*

Temporal Stability. Whereas internal consistency reliability is concerned with the relatedness of each item on a test (Cho & Kim, 2015), ***temporal stability/reliability*** is concerned with the consistency of test scores across repeated testings (Allen & Yen, 1979). Estimating temporal stability presumes that the focal construct should be relatively stable/invariant over the time frame being examined. There is not a single, recommended time lag to use in a test-retest design. As Crocker and Algina (1986) noted, "There is no single answer. The time period should be long enough to allow effects of memory or practice to fade but not so long as to allow maturational or historical changes to occur in the examinees' true scores" (p. 133). In some instances, a lag of two or three weeks may be sufficient, in other instances the researcher may decide to implement a longer time lag.

Historically, temporal stability has been estimated by administering the same test to the same group of examinees at different points in time. However, DeSimone (2015) identified several problems with this strategy. First, this approach is effectively ignoring item-level psychometric information. Because the test-retest correlation is a function of the total test score (i.e., composite of all items), it could be masking individual items that may be problematic. Second, this approach to estimating reliability may be considered a form of the logical fallacy called "affirming the consequent" (see DeSimone, 2015, p. 135). Specifically, a high estimate of temporal stability could be obtained because the item-level relationships are consistent across time or the estimate could be engendered by different response patterns that result in a similar total score. To illustrate this problem, DeSimone (2015, p. 135) used the following example:

> *If a respondent's item responses on a sum-scored, five-item, five-option, Likert-based questionnaire are 2, 2, 3, 5, 4 at Administration 1 (A1) and 5, 4, 4, 1, 2 at Administration 2 (A2), the scale score remains identical (16) across both administrations.*

Thus, to rectify both of these problems, DeSimone (2015) recommends examining both the scale-level and item-level patterns of stability. At the item level, researchers may compute a variant on the standardized root mean-square residual (SRMR) focused on temporal consistency (*TC*):

$$SRMR_{TC} = \sqrt{\sum_{g=1}^{K} \frac{(r_{g1} - r_{g2})^2}{K}} \quad (23),$$

where r_{g1} and r_{g2} refer to the g=1 to *K* diagonal item-level diagonal correlations at time 1 and time 2. Estimates may range "from zero to one, with lower values indicating more similarity between inter-item correlation matrices" (p.135, DeSimone, 2015). Alternatively, researchers could compare the temporal stability of items and test scores using methods that have traditionally been applied to tests of factor invariance or equivalence (see DeSimone, 2015 for a more detailed discussion).

- *Researchers interested in obtaining estimates of temporal stability should compute estimates at both the scale and item level (e.g., SRMRTC; DeSimone, 2015) and/or examine temporal stability using tests of measurement invariance/equivalence.*

Finally, DeSimone (2015) discusses a new statistic developed to identify respondent-level temporal inconsistency, denoted D_{ptc}. Importantly, this new statistic can account for examinees

who put forth insufficient effort when responding to tests (e.g., provide the same answer to multiple items without reading the items), even when tests do not include items specifically designed to capture this phenomenon.

- *Researchers are encouraged to estimate Dptc when there are concerns examinees may have exerted insufficient effort during the testing process.*

Standards for reliability. Many researchers invoke a minimum threshold for acceptable reliability of .70 and reference Nunnally (1978) as support for using this threshold. However, as Lance, Butts, and Michels (2006) noted, this threshold and reference represent a form of methodological urban myth that, unfortunately, has been perpetuated over the last 40 years. Instead, when discussing ***standards for reliability,*** Nunnally's (1978) actually stated:

> "In the early stages of research…one saves time and energy by working with instruments that have only modest reliability, for which purpose reliabilities of .70 or higher will suffice…In contrast to the standards in basic research, in many applied settings a reliability of .80 is not nearly high enough. In basic research, the concern is with the size of correlations and with the differences in means for different experimental treatments, for which purposes a reliability of .80 for the different measures is adequate. In many applied problems, a great deal hinges on the exact score made by a person on a test…In such instances it is frightening to think that any measurement error is permitted. Even with a reliability of .90, the standard error of measurement is almost one-third as large as the standard deviation of the test scores. In those applied settings where important decisions are made with respect to specific test scores, a reliability of .90 is the minimum that should be tolerated, and a reliability of .95 should be considered the desirable standard." (pp. 245-246)

This guidance was reiterated in Nunnally and Bernstein (1994) (see. p. 265) and is consistent with other recommendations concerning the use of tests and measures that will be used in practice. For example:

- "[desirable reliability coefficients] usually fall in the .80s or .90s" (p. 78; Anastasi, 1968).
- "If a procedure is to be used to compare one individual with another, reliability should be above .90" (p. 145; Cascio & Aguinis, 2019).
- "Following the leadership of T. L. Kelley there has been general agreement that to be sufficiently reliable for discriminating very accurately between individuals, a test should have a minimum reliability coefficient of at least .94. Some have been more liberal in this regard, allowing a minimum of .90." (Guilford & Fruchter, 1973, p. 91).
- "the minimum acceptable level of reliability for psychological measures in the early stages of development is .70 (Nunnally, 1978). Higher levels may be required of measures . . . used in advanced field research and practice." (LeBreton & Senter, 2008, p. 839)

- *Researchers who will be using test scores to draw inferences about specific individuals in applied settings (e.g., whom to hire, fire, promote, reward, or punish) should strive to*

draw those inferences from highly reliable tests-estimates of reliability should be .90 or higher.

- *Researchers who will be using test scores to draw general inferences about group differences or who will be drawing inferences from tests that are in the early stages of development (i.e., basic research) should strive to draw those inferences from tests with at least moderate levels of reliability-estimates should exceed .65.*

Factor analysis. Factor analysis is a term used to refer to a broad set of statistical procedures that are applied to the correlations or covariances between variables (i.e., items or preferably, between tests or subtest comprised of multiple items). The purpose of factor analysis is to determine whether the covariances between the set of variables may be represented using a smaller number of latent variables or dimensions, denoted ***factors.*** Essentially, factor analysis involves decomposing the observed covariance matrix into component matrices representing different sources of variance: error (co)variance and factor (i.e., true score) (co)variance. The fundamental equation of factor analysis is given as follows:

$$\Sigma = \Lambda\Phi\Lambda' + \Theta \quad (24),$$

where, Σ is a K by K covariance or correlation matrix for the observed items, Λ is a K by P matrix of factor loadings where P < K, Φ is a P by P covariance or correlation matrix for the P latent factors, Λ' is simply the transpose of the original Λ matrix, and Θ is a K by K matrix with error variances on the major diagonal and zeros in the off-diagonal elements (Long, 1983; Mulaik, 2009). Although a number of different heuristics have been offered for guiding the interpretation of factor/component loadings, Tabachnick and Fidell (2013) provided a nice discussion of interpreting factor loadings:

> "As a rule of thumb, only variables with loadings of .32 and above are interpreted. The greater the loading, the more the variable is a pure measure of the factor. Comery and Lee (1992) suggest that loadings in excess of .71 (50% overlapping variance) are considered excellent, .63 (40% overlapping variance) very good, .55 (30% overlapping variance) good, .45 (20% overlapping variance) fair, and .32 (10% overlapping variance) poor. Choice of the cutoff for size of loading to be interpreted is a matter of researcher preference" (p. 654).

If the items being analyzed were designed to measure a single factor, then the Φ matrix is expected to drop from the equation-but this can be tested empirically by extracting different numbers of factors and comparing the fit of single factor and multifactor models. If the items being analyzed were designed to measure multiple factors, then the Φ matrix is likely to be retained. In addition, when multiple factors are measured, one hopes that the estimated Λ matrix conforms to a pattern known as simple structure (i.e., when each item has a strong loading on a single factor a zero or near zero loadings on all remaining factors).

Exploratory factor analysis. In the early stages of test development and validation, it is common for researchers to conduct an ***exploratory factor analysis (EFA)*** on the items comprising the test. An EFA is so named because it contains minimum constraints and thus involves estimating nearly all of the elements comprising the right-hand side of equation 24. The

only constraints placed on the model are the number of factors extracted and diagonal structure of the error matrix, Θ. EFA is typically used in the early stages of scale development and validation by researchers striving to identify items that fail to load on the appropriate factor, or that might load on multiple factors (Allen & Yen, 1979; Hinkin, 1998).

Fabrigar, Wegener, MacCallum, and Strahan (1999) identified five methodological decisions or issues that must be addressed by researchers opting to use EFA. Below these issues are summarized, and where appropriate, additional guidance is offered.

1. *Researchers must determine the variables to include and the sample to analyze.*

The reader is directed to earlier portions of this report discussing variable and sample selection. In addition, Velicer and Fava (1998) provide excellent suggestions concerning the impact of variable and subject sampling on the accuracy of EFA results.

2. *Researchers must determine whether EFA is really the most appropriate analysis.*

Fabrigar et al. (1999) noted that researchers sometime incorrectly use EFA when other procedures might be more appropriate. Recall that the purpose of EFA is to discover the number of latent factors underlying an observed covariance matrix and to estimate the pattern of factor loadings. In doing so, researchers are functionally partitioning the observed covariance matrix, Σ, into a portion engendered by a (common) set of latent constructs, ΛΦΛ^', and a portion that may be attributed to (unique) item-specific measurement error, Θ.

In contrast, ***principal component analysis (PCA)*** does not distinguish between common (construct) and unique (error) sources of variance. Instead, this approach strives to create a set of observed "components" that are mathematically defined as a weighted linear combination of the observed items. Thus, PCA is optimally suited for creating variance maximizing weighted linear composites (Fabrigar et al., 1999; Tatsuoka & Lohnes, 1988). Thus, researchers interested in (exploring or testing) the factor structure of a measure are encouraged to stay away from PCA and instead explore whether whether EFA or a *confirmatory factor analysis (CFA)* may be more appropriate. CFA is briefly discussed in the next section of the report.

3. *Researchers must determine the specific procedures that will be used to fit the model.*

A number of different methods exist for extracting latent factors from the observed covariance matrix, including: principal axis with prior estimation of communalities, iterative principal axis, maximum likelihood, alpha, minimum residual, image, generalized least squares, and unweighted least squares. Each of these methods seeks to extract factors by minimizing or maximizing some target function. For example, alpha factoring extracts factors with the goal of maximizing coefficient alpha and unweighted least squares strives to minimize the squared differences between the original sample covariance matrix and the covariance matrix that is estimated (i.e., reproduced) after estimating the elements on the right-hand side of equation 24. More detailed discussions of these methods are available in Fabrigar et al. (1999), Tabachnick and Fidell (2013), Mulaik (2009), Tatsuoka and Lohnes (1988), and Velicer and Fava (1998).

4. *Researchers must determine how many factors to extract. Although multiple approaches exist, parallel analysis likely offers the most accurate results.*

Prior to using the above methods to extract factors and compute factor loadings, researchers must first determine how many factors they wish to extract. Multiple methods exist including: scree plots, Kaiser's criterion, and parallel analysis. Although each of these techniques have their strengths and limitations, there is growing consensus that parallel analysis (Horn, 1965) provides the most accurate conclusions regarding the number of factors to retain. Hayton, Allen, and Scarpello (2004) provided a brief review of the criteria used to determine the number factors and offer a step-by-step guide to undertaking a parallel analysis.

5. *Determining the method used to rotate the initial factor solution.*

After determining how many factors to extract and how to estimate the loadings of the items onto those factors, researchers may also wish to rotate the initial factor solution. Rotations are a tool that researchers can use to help improve the interpretability of the factor analysis. Functionally, rotations involve redefining the latent factor, and thus the relationship between the latent factor and the observed items. Thus, rotation serves to change the magnitude and pattern of the initial factor loadings, with the goal of improving the interpretation of the loading matrix. Rotations may allow the latent factors to correlate (oblique) or constrain them to be uncorrelated (orthogonal; see Mulaik, 2009; for a less technical treatment see Tabachnick & Fidell, 2013). For example, the first unrotated component from a PCA applied to a set of cognitive tests (e.g., ASVAB scores) will often yield evidence for a single dominant component – sometimes denoted, "g" to represent general mental ability. However, rotating the original solution is likely to result in an alternative representation of the data. For example, applying rotations to a set of cognitive tests is likely to reveal clusters of tests that are designed to measure common attributes (e.g., verbal ability, quantitative ability, spacial ability).

Confirmatory factor analysis. As noted earlier, researchers may also opt to estimate a ***confirmatory factor analysis*** or *CFA.* A CFA is similar to an EFA, but rather than allowing the computer to "explore" the data to determine the number of latent factors and the pattern of factor loadings, researchers impose a series of constraints on the elements comprising equation 24. Thus, researchers create an a priori model based on a set of restrictions or constrains applied to equation 24 and then test (i.e., confirm or disconfirm) the fit of that model to the data. Constraints that may be imposed included: number of factors to extract, pattern of factor loadings, equivalence (or lack thereof) of factor loadings or error variances, and pattern of inter-factor correlations (see Long, 1983). Because of the degree of specificity needed when identifying the constraints for a CFA, this procedure is typically invoked later in the scale development and validation process (Fabrigar et al., 1999)-after researchers have a better conceptual and empirical understanding of how their items are related to the latent construct(s).

Statistically, a CFA involves estimating the elements of equation 24 after the researcher has imposed the necessary constraints on Λ, Φ, and Θ. Those matrices are then used to compute a reproduced (or estimated) covariance (or correlation) matrix, $\Sigma\hat{}$. This matrix is then compared to the original covariance (or correlation) matrix to determine whether the constrained model is consistent with the data. If there is a strong "fit" between the reproduced matrix and the original matrix, then one may conclude that the constrained model is consistent with the data. It is important to recognize that multiple models may engender similar levels of "fit." Thus, if there is strong fit between the model and the data, one can only infer that the model is consistent with the data, not that the model is proven to be the one, correct model. A number of fit statistics are

available to researchers. For a detailed review and critique, the interested reader is directed to: Benter (1990), Browne and Cudeck (1993), Hayakawa (2018), Mulaik, James, Van Alstine, Bennett, Lind, and Stilwell (1989), Widaman and Thompson (2003); for a discussion surrounding the accurate interpretation of model fit statistics see Lance et al. (2006).

- *Researchers should report the multiple goodness-of-fit indexes when using CFA.*

- *Researchers should report corrected goodness-of-fit indexes when the number of manifest indicators (i.e., items/tests) is large relative to the sample size (see Hayakawa, 2019).*

1.5 Using Item Response Theory to Evaluate Items and Build Tests

1.5.1. Overview of Item Response Theory

1.5.1.1 Typical Assumptions of Item Response Theory

Unidimesionality. Most Overview of Item Response Theory (IRT) models are predicated on the assumption that a set of items is designed to measure a single latent construct (i.e., the correlations between items may be accounted for by a single construct; Crocker & Algina, 1986). This assumption of ***unidimensionality*** is rarely met in practice because responses to any given set of items is likely to be influenced by a host of secondary constructs including cognitive, personality, and test-taking factors. However, Hambleton, Swaminathan, and Rogers (1991) suggested that one may conclude that the unidiminsionality assumption has been satisfied when there is "the presence of a "dominant" component or factor that influences test performance" (p. 9). Although it is important to test for unidimensionality, researchers have concluded that "IRT model parameter estimation is fairly robust to minor violations of unidimensionality, especially if the latent-trait dimensions (factors) are highly correlated or if secondary dimensions are relatively small" (p. 231; Embretson & Reise, 2000).

- *Researchers should test the assumption of unidimensionality to ensure that there is a "dominant" latent factor that appears to strongly influence test performance.*

    - *Undimensionality may be tested using EFA, CFA, parallel analysis, modified parallel analysis, (see prior sections on EFA/CFA for references) or categorical principal components analysis (Mair, 2018).*

Local independence. IRT models also assume that when the constructs "influencing test performance are held constant, examinees' responses to any pair of items are statistically independent" (Hambleton et al., 1991, p. 10). This assumption of ***local independence*** essentially states that *after* accounting for the latent construct of interest, there should be no residual covariance remaining between pairs of items. If one's data are unidimensional, then, by definition, they are locally independent and these concepts become one in the same (Lord & Novick, 1968). However, it is possible to achieve local independence when data are multidimensional, given that the complete set of latent constructs influencing test performance has been identified/specified (de Ayala, 2009; Hambleton et al., 1991).

Several authors have suggested that Yen's (1993) Q3 statistic is a reasonable statistic to use when testing for local independence (de Ayala, 2009; Embretson & Reise, 2000; This statistic is

essentially "the correlation between the residuals for pairs of items" (de Ayala, 2009, p. 132). It is computed by first estimating the item-level residuals by subtracting the estimated item responses from the actual item responses:

$$d_{ig} = X_{ig} - p_g(\hat{\theta}_i) \quad (25),$$

$$d_{ih} = X_{ih} - p_h(\hat{\theta}_i) \quad (26),$$

where, $d_{ig}$and d_{ih} refer to the residuals for person *i* on obtained difference observed scores on items *g* and *h* from the predicted scores, $p_g(\hat{\theta}_i)$ and $p_h(\hat{\theta}_i)$. The Q3 statistic is estimated by pairwise correlating these residuals across the $g = 1$ to K items on the test:

$$Q_{3_{gh}=r_{d_g d_h}} \quad (27).$$

As Embretson and Reise (2000) noted, "the expected value of Q_3 under the hypothesis of local independence is -1/(N-1). Thus, in large samples a researcher would expect Q_3 to be around zero and large positive values indicate item pairs that share some other factor that may be a cause for concern" (p. 232). Chen and Thissen (1997) found that the Q3 statistic was more powerful to detect underlying local dependence compared to several alternative statistics, including the Local Dependence square ($LD - \chi^2$), and was equally powerful for detecting surface local dependence. Item pairs with $LD - \chi^2 > 10$ should be examined for possible violations of the local independence assumption (Cole & Paek, 2020).

- *Researchers are encouraged to formally test the assumption of local independence using statistics such as Q3 or $LD - \chi^2$.*

Item characteristic curves. As Embretson and Reise (2000) noted, in addition to the assumption of local independence, IRT models are predicated on the assumption that "the item characteristic curves have a specified form" (p. 45). The item characteristic curve (or ICC) may be thought of as the "basic building block of item response theory" (Baker & Kim, 2017, p. 3). It represents the foundation upon which all aspects of IRT are built. ICCs provide a graphical depiction of the relationship between the latent construct, denoted θ, and the probability of selecting a particular response alternative on the focal item, p_g. Whereas CTT assumes that the relationship between an item and the latent construct is linear, IRT relaxes this assumption to allow for non-linear relationships between items and the latent construct. Visually, ICCs may be represented as the nonlinear regression of the probability that a response option is selected (p_g) onto the latent trait (θ; which is typically assumed to have $M = 0$ and $sd = 1$). Different IRT models invoke distinct assumptions about the nature of the item-construct relationship. As a result, different models yield different ICCs. However, irrespective of the particular model selected, all ICCs are a function of at least two item parameters: ***item difficulty*** and ***item discrimination.*** Although these labels were used in CTT, they are defined differently in IRT.

1.5.1.2 Item Difficulty

In IRT, ***item difficulty*** is defined as the level of the latent trait where the probability of endorsing the correct or keyed item response is .50 (Hambelton, Swaminathn, & Rogers, 1991) and for any

given focal item, X_g, this parameter may be denoted b_g. Thus, item difficulty serves as a "location" parameter that shifts the ICC left or right along the construct continuum (Embretson & Reise, 2000).

1.5.1.3 Item Discrimination

In IRT, ***item discrimination*** provides information about the steepness of the ICC when $\theta = b_g$. This value is denoted a_g and is not exactly equal to the slope, but is proportional to it (Hambleton et al., 1991). The steeper the slope, the more effective the item is at discriminating between examinees falling at different levels of the construct continuum. Here, researchers are looking for items with positive item discrimination parameters with larger values indicating that the item is more effective at discriminating between individuals falling on different levels of the latent construct. Negative item-discrimination parameters, are typically interpreted as indicating that the item is miskeyed or simply a problematic item. Negative values indicate that the probability of endorsing the keyed item response decreases as levels of the latent construct increase. For example, if an item comprising the ASVAB had a negative item discrimination parameter, it would indicate that individuals with higher levels of cognitive ability were more likely to endorse the incorrect item response. Bear in mind that slopes represent discrimination. The item difficulty represents the location on the construct continuum. So, researchers would want to vary item difficulty levels if they were building a general test designed to tell them about a wide range of scores. In contrast, if they were only interested in maximizing the discrimination of a single point on the construct continuum, then they would only want to select items with difficulty levels near that level of the construct.

1.5.1.4 Examples of Item Characteristic Curves

To illustrate how item difficulty and item discrimination parameters impact ICCs, several illustrative example ICCs were computed and are described below. Figure 3 contains the ICCs for three items with a common (i.e., fixed) level of item difficulty, $b_g = 0$, but with varying levels of item discrimination. The common item difficulty indicates that all items have a 50% chance of being correctly answered by individuals with a latent trait score of $\theta = 0$. However, these items differ in the degree to which they are effective at discriminating between levels of the latent construct. The flat line represents an ICC where the item discrimination parameter was set to 0. The two remaining ICCs were estimating using discrimination parameters set to values of 1 and 2. As the magnitude of the item discrimination parameter increases, so too does the slope of the curve. Thus, ceteris paribus, an item with a discrimination parameter of 2 does a better job of distinguishing between levels of the latent construct compared to an item with a discrimination parameter of 1. Figure 4 contains the ICCs for three items with a common discrimination parameter, $a_g = 1$, but with varying levels of item difficulty of -2, 0, and +1.5. Thus, each of these items provides similar levels of discrimination between levels of the construct, but they differ in terms of where this discrimination is going to be optimized. The item with a difficulty of -2.0 provides the greatest discrimination at low levels of the construct. In contrast, the item with a difficulty of 0 provides maximum discrimination in the middle of the construct continuum (i.e., at the mean).

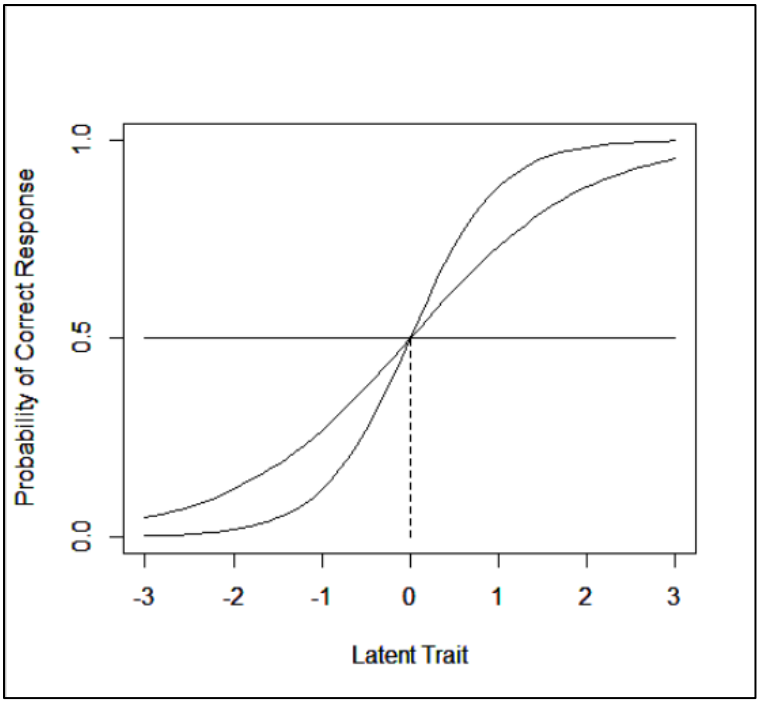


Figure 3. Three ICCs with Fixed Item Difficulty and Varying Levels of Item Discrimination

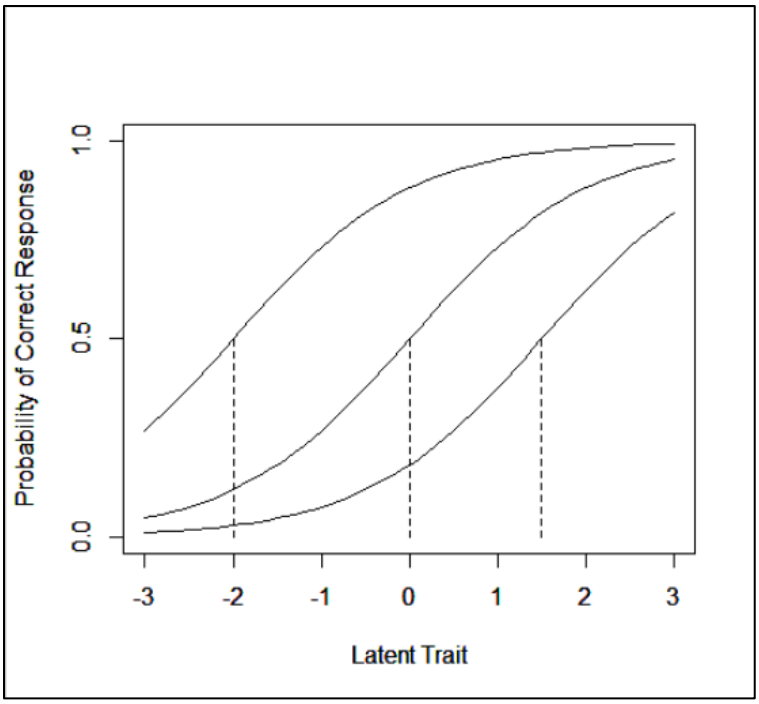


Figure 4. Illustrative ICCs with Varying Levels of Item Difficulty and Fixed Item Discrimination

1.5.2. Item Response Models for Dichotomously Scored Items

1.5.2.1 1-Parameter Logistic Model

The simplest IRT model for the analysis of dichotomously scored items is the ***1-parameter logistic model*** (1PL), which is so named because it only estimates a single item parameter-item difficulty. ICCs for the 1PL are based on:

$$P\left(X_{gi}=1|\theta_i,b_g\right)=\frac{e^{(\theta_i-b_g)}}{1+e^{(\theta_i-b_g)}} \quad (28)$$

where,

$P\left(X_{gi}=1|\theta_i,b_g\right)$ = conditional probability that examinee *i* will endorse the correct or keyed response on item *g*, b_g = the difficulty parameter for item *g, e* = base of the natural logarithm θ_i = score on the latent trait for examinee *i*. The distribution of θ_i is typically assumed to have a mean of 0 and standard deviation of 1.

Although not apparent in equation 28, there is a constant item discrimination parameter, *a*, which is fixed to a value of 1 across all items. An alternative presentation of the 1PL model includes an estimated, but fixed item discrimination parameter (i.e., *a* is estimated rather than constrained to unity, but the estimated value is applied to all items). The absence of item-specific subscripts indicate the *a* parameter is fixed to a common value across each of the *g* items:

$$P\left(X_{gi}=1|\theta_i,b_g\right)=\frac{e^{(a(\theta_i-b_g))}}{1+e^{(a(\theta_i-b_g))}} \quad (29).$$

Some researchers distinguish between equations 28 and 29 using the labels of the Rasch model and the 1PL model, respectively. Although the models may be thought of as statistically equivalent, there are some conceptual differences between these models. The interested reader is directed to pp. 11-19 in de Ayala (2009).

1.5.2.2 2-Parameter Logistic Model

One of the most popular models for the analysis of dichotomously scored items is the ***2-parameter logistic model*** (2PL), which is so named because it yields estimates of two item parameters-item difficulty and item discrimination. ICCs for the 2PL are based on:

$$P\left(X_{gi}=1|\theta_i,b_g,a_g\right)=\frac{e^{(a_g(\theta_i-b_g))}}{1+e^{(a_g(\theta_i-b_g))}} \quad (30),$$

where, $P\left(X_{gi}=1|\theta_i,b_g,a_g\right)$ = conditional probability that examinee *i* will endorse the correct or keyed response on item *g*, b_g = the difficulty parameter for item *g,* a_g = the discrimination parameter for item *g,* e = base of the natural logarithm θ_i = score on the latent trait for examinee *i*. The distribution of θ_i scores is typically assumed to have a mean of 0 and standard deviation of 1. Because each item is allowed to have a unique discrimination parameter, the steepness of the ICCs is allowed to vary across items.

1.5.2.3 3-Parameter Logistic Model

Another popular model for analyzing dichotomously scored items is the ***3-parameter logistic model*** (3PL), which is so named because it yields estimates of three item parameters-item difficulty, item discrimination, and a lower asymptote or guessing parameter. ICCs for the 3PL are based on:

$$P\left(X_{gi} = 1|\theta_i, b_g, a_g, c_g\right) = c_g + (1 - c_g)\frac{e^{(a_g(\theta_i - b_g))}}{1+e^{(a_g(\theta_i - b_g))}} \quad (31),$$

where, $P\left(X_{gi} = 1|\theta_i, b_g, a_g, c_g\right)$ = conditional probability that examinee *i* will endorse the correct or keyed response on item *g*, b_g = the difficulty parameter for item *g*, a_g = the discrimination parameter for item *g*, c_g = the lower-asymptote or guessing parameter for item *g*, *e* = base of the natural logarithm, and θ_i = score on the latent trait for examinee *i*. Again, the distribution of θ_i scores is typically assumed to have a mean of 0 and standard deviation of 1.

Like ICCs based on the 2PL, the ICCs derived from the 3PL allow each item to assume a unique difficulty and discrimination parameter. In addition, the 3PL also allows researchers to adjust the floor for the range of conditional probabilities by allowing each item to assume a unique guessing parameter. These guessing parameters provide an estimate of the likelihood of endorsing the correct (or keyed) response option, even for examinees with extremely low trait levels. For example, if a researcher is using a test that combines multiple choice questions with four response alternatives (e.g., A, B, C, D) with multiple choice questions having only two response alternatives (e.g., True, False), he or she will likely want to adjust the range of lower guessing parameters. In the first instance, respondents have a 25% chance of correctly guessing the answer; whereas, in the second instance, respondents have a 50% chance of correctly guessing the answer.

1.5.2.4 4-Parameter Logistic Model

Another model that may be used with dichotomously scored items is the 4-parameter logistic model (4PL). Whereas the 3PL model includes a lower-bound guessing parameter for low ability examinees (i.e., a lower asymptote parameter), the 4PL model adds an upper-bound slip parameter designed to "accommodate high ability examinees' mistakes (incorrect answers) due to their carelessness or some other reasons" (Paek & Cole, 2020, p. 88):

$$P\left(X_{gi} = 1|\theta_i, b_g, a_g, c_g, d_g\right) = c_g + (d_g - c_g)\frac{e^{(a_g(\theta_i - b_g))}}{1+e^{(a_g(\theta_i - b_g))}} \quad (32).$$

The item response function for the 4PL model is essentially the 3PL model but instead of setting the upper asymptote value to 1, it is estimated as the slip parameter-d_g (see also Loken & Rulison, 2010).

1.5.2.5 Other Models

Polytomous item response models. A number of IRT models are available for modeling items with more than two response categories (i.e., polytomously scored items) including: the partial credit model (Masters, 1982), the rating scale model (Andrich, 1979), the generalized

partial credit model (Muraki, 1992), the graded response model (Samejima, 2010), and the nominal response model (Bock, 1972). Due to space constraints, these models are not reviewed in the current report. The interested reader is directed to original citations noted above, as well as, more recent treatments of these models by: de Ayala (2009), Baker and Kim (2017), Embretson and Reise (2000), Mair (2018), or Paek and Cole (2020). The latter book includes the R code used to calibrate these models.

Multidimensional item response models. As the name implies, multidimensional IRT models are used when a researcher believes that more than one latent construct is needed to describe examinees' item responses (de Ayala, 2009; Paek & Cole, 2020). Basically, these models assume that the latent construct space is multidimensional, or as de Ayala (2009) summarized, "…in some situations it may be more realistic to hypothesize that a person's response to an item is due to his or her locations on multiple latent variables" (p. 275).

Paek and Cole (2020) noted that there are two general families of multidimensional IRT models-between-item and within-item models. Between-item models are appropriate when each item on a test is thought to be influenced by one of several different latent traits. Between-item models are sometimes referred to as simple structure models because items are expected to form distinct clusters or factors reflecting the different constructs. In contrast, within-item multidimensional IRT models are appropriate when responses to the item may be simultaneously influenced by multiple latent traits. As such, Paek and Cole (2020) suggested that within-item models might be conceptualized as having a "cross-loading item structure" (p. 198).

A large number of multidimensional IRT models are available. Paek and Cole (2020) provide illustrative examples using models based on between-item multidimensionality. Specifically, these authors provide the R code for calibrating multidimensional extensions of the 1PL, 2PL, 3PL, partial credit model, generalized partial credit model, and the grade response model. The authors also provide an illustrative example using a multidimensional 2PL model to calibrate items based on within-item multidimensionality. The interested reader is directed to de Ayala (2009), Mair (2018), and Paek and Cole (2020).

1.5.3. Evaluating Items and Building Tests

1.5.3.1 Standard Error of Estimate and Item information

Standard error of estimate. Under IRT models, researchers are able to generate estimates of examinees' true scores-that is, their standing on the latent construct. Any sample estimate of a person's location on the construct continuum will be subject to error. As de Ayala (2009) noted, "…in IRT our uncertainty about a person's location can be quantified through the estimate's ***standard error of estimate*** (SEE), $\sigma_e(\hat{\theta})$ …the SEE specifies the accuracy of $\hat{\theta}$ with respect to the person location parameter, θ" (p.27; italics & bold type added). The SEE is quantitatively and qualitatively different from the standard error of measurement from CTT. The latter assumes a single, fixed value that is applied to all persons across all levels of the construct continuum. In contrast, the SEE varies across different levels of θ. Thus, it is possible to have very accurate estimates of a person's trait level for some levels of θ and less accurate estimates for other levels of θ (de Ayala, 2009; Embretson & Reise, 2000; Hambleton et al., 1991).

Item and test information. Rather than ask about how the degree of variability or error in our estimates of a person's location on the trait continuum changes as we move up or down the continuum, we could ask "how much information do we have about a person's location [on the trait continuum]?" (de Alaya, 2009, p. 29). The concept of ***item information*** is best understood by examining the equation used to estimate item information under the 3PL model:

$$I_g(\theta) = \frac{2.89a_g^2(1-c_g)}{\left[c_g+e^{1.7a_g(\theta-b_g)}\right][1+e^{-1.7a_g(\theta-b_g)}]^2} \quad (33).$$

As Hambleton et al. (1991) noted, "…it is relatively easy to infer the role of the *b, a,* and *c* parameters in the item information function: (a) information is higher when the *b* value is close to θ than when the *b* value is far from θ, (b) information is generally higher when the *a* parameter is high, and (c) information increases as the *c* parameter goes to zero" (p. 91). The information function for a set of items (i.e., a test) is simply a unit-weighted sum of the item information functions:

$$I(\theta) = \sum_{g=1}^{K} I_g(\theta) \quad (34).$$

It is important to note that these values are conditional on the level of the latent trait, θ. Thus, a set of items will furnish different levels of information at different levels of the latent trait. Figure 5 graphs the item information curves for four items with a common discrimination parameter of 1.5 and difficulty parameters with values of -2.0, 0.0, 0.5, and 1.5. From this figure it is possible to see the small "humps" in the levels of information correspond exactly to the location of the item on the theta continuum (i.e., difficulty parameters). Note that if you wanted to build a test that was designed to measure a wide range of theta levels, then you would want to include items with varying difficulty levels (i.e., b parameters), because information is maximized when the theta level is close to the b parameter. So, if you want to maximize information across the range of theta, you would need to include items that varying in their difficulty/location on theta.

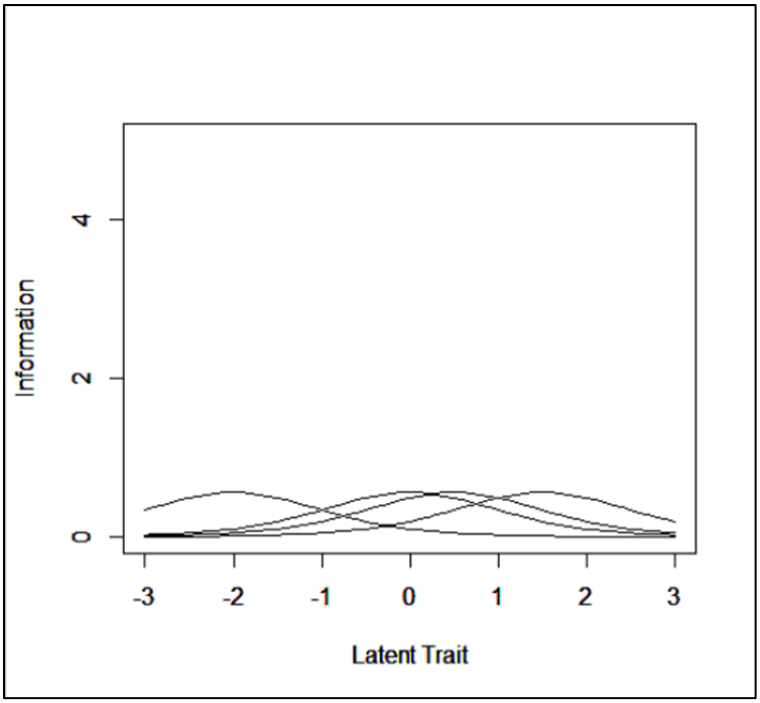


Figure 5. Item Information Curves for Items with Fixed Discrimination and Varying Levels of Difficulty

Figure 6 illustrates the inverse relationship between test information and the SEE. This relationship is given by (Embretson & Reise, 2000; Hambelton et al., 1991):

$$SE(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}} \quad (35).$$

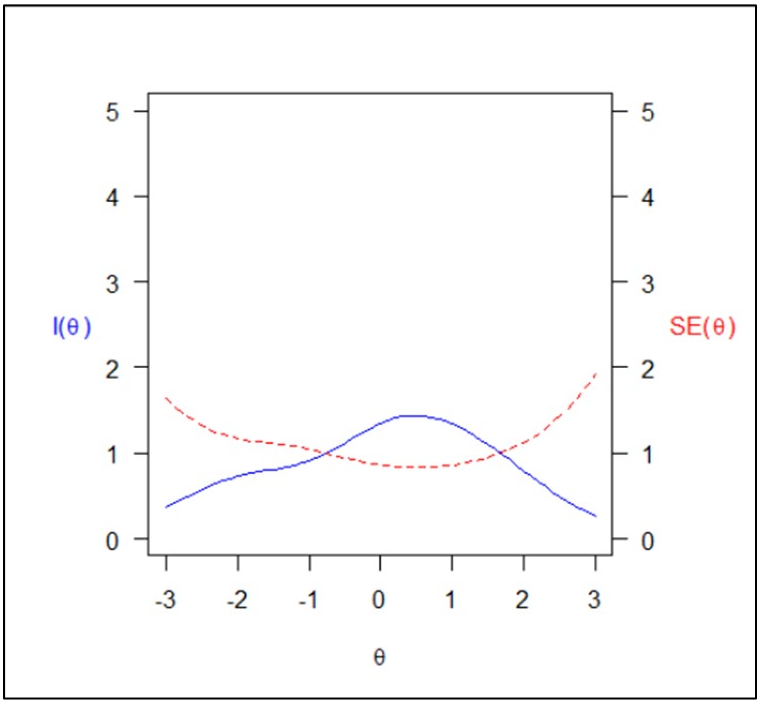


Figure 6. Relationship between Test Information Curve and Standard Error of Estimate

Target information function. Prior to calibrating IRT models, researchers should first articulate the purpose of testing and develop a target information function (Lord, 1977) consistent with that purpose and that will guide item evaluations using item information curves.

- *If the purpose of testing is to generally discriminate across all levels of the construct continuum, then researchers should specify a uniform (i.e., quasi-rectangular) target information function.*

- *If the purpose of testing is a fine-grained distinction at a specific level of the construct (e.g., at a particular cut-point), then researchers should specify a target information function that maximize information around the desired cut-point (i.e., a peaked) target information function)*

1.5.3.2 Item Parameters

As illustrated in equation 33, item information is a direct function of the item parameters. Thus, decisions concerning which items should be included as part of a test will depend largely on the target information function that researchers are trying to approximate. This is especially the case for item difficulty parameters. For example, if a researcher is seeking to maximize information at +2 θ (e.g., because this level of θ was determined to represent an important cut-point or threshold), then the items selected will differ appreciably from those that would be selected if the goal of the researcher was to have similar levels of information across all levels of θ. In contrast, Baker and Kim (2017) provided heuristics for interpreting item discrimination (see Table 8). Some additional (general) recommendations for interpreting item parameters include:

- *Item difficulty: researchers should select items with a range of b_g values if the goal is to develop a test that provides information across the entire continuum; if researchers wish to maximize information at a particular point on the trait continuum (e.g., +2 θ), then they should select items with b_g values close to this level of θ.*

- *Item discrimination: researchers should select items with positive discrimination values. Items with negative discrimination should be examined to determine if a coding error has occurred; if no error has occurred, the item should be removed. Ceteris paribus, larger values for item discrimination are preferred over smaller values.*

- *Item guessing: Ceteris paribus, items with smaller values for the guessing parameter are preferred over items with larger values.*

Table 8. Standards for Interpreting Item Discrimination Parameters

Verbal label	Range of values	Typical values
None	0	0.00
Very low	0.01-0.34	.18
Low	0.35-0.64	.50
Moderate	0.65-1.34	1.00
High	1.35-1.69	1.50
Very high	>1.70	2.00
Perfect	+∞	+∞

Note. Reproduced from "The basics of item response theory using R," F. B. Baker & S. Kim, 2017, Table 2.4, p. 26, Copyright 2017 by Springer International Publishing.

1.5.3.3 Omnibus Tests of Model-Data Fit

An additional consideration when selecting items for retention as part of a test is whether the specified IRT model is consistent with the data-that is, the extent which there is evidence of ***model-data fit.*** Embretson and Reise (2000) and de Ayala (2009) reviewed a number of different statistics used to test the degree of model-data fit.

Likelihood ratio test. The first statistic "is based on the likelihood ratio (G2) test statistic for comparing the relative fit of hierarchically nested models" (de Ayala, 2009, p. 140). For example, a researcher could use a 2PL model to calibrate a 15-item survey and then compare the fit of that model to one obtained using the 3PL model. The relative fit would be tested using:

$$\Delta G^2 = -2\ln(L_R) - \left(-2\ln(L_F)\right) = G_R^2 - G_F^2 \quad (36),$$

where, *LR* is the maximum of the likelihood function for the restricted model and LF is the maximum of the likelihood for the unrestricted or full model. The resulting difference in likelihood ratios is distributed as a chi-square statistic with degrees of freedom equal to the difference in the number of parameters estimated in the restricted and unrestricted models. Returning to our example, the restricted model would be the 2PL model because all item guessing parameters are essentially constrained to zero in this model. (Note that the 1PL model would be more restricted as it forces all items to assume a common slope.) The unrestricted model would be the 3PL model because item guessing parameters are now being freely estimated. Thus, the difference in likelihood ratios obtained using equation 36 would be tested against the critical value for a chi square with 15 degrees of freedom (i.e., 24.996 for $p < .05$).

Change in variance explained. The second model-data fit statistic is analogous to testing changes in R^2 using hierarchical regression analysis (de Ayala, 2009). This approach is essentially examining whether the unrestricted model accounts for more variance relative to the restricted model:

$$R_\Delta^2 = \frac{(G_R^2 - G_F^2)}{G_R^2} \quad (37).$$

Because the unrestricted model will almost always have better fit to the data, one should remember to interpret this statistic through the lens of an effect size (i.e., the proportion of additional variance that is accounted for by using the more complex model).

Information criteria. Given that the unrestricted models will tend to have better fit than the restricted models, some researchers have advocated for the third type of model-data fit statistic-information criterion measures-that adjust model fit estimates by considering the complexity of the model. de Ayala (2009) reviewed two statistics in this information criterion tradition: the Bayesian information criterion (BIC) and the Akaike information criterion (AIC) and researchers typically report either or both of them. These statistics essentially correct the log likelihood estimates of model-data fit:

$$BIC = -2lnL + \ln(N) * Nparm \quad (38), \text{ and}$$

$$AIC = -2lnL + 2 * Nparm \quad (39),$$

where, *Nparm* refers to the total number of parameters estimated by the model and *N* refers to the number of participants in the sample. BIC and AIC are both interpreted such that, smaller values indicated better model-data fit.

M_2. A final test of omnibus model-data fit is provided by the limited information goodness of fit statistic introduced by Maydeu-Olivares and Joe (2006) and denoted, ***M_2***. As de

Ayala (2009) noted, this "statistic maintains appropriate Type I error rates under varying degrees of model misspecification" and is also distributed as a chi-square statistic. This statistic is growing in popularity and is being recommended with greater frequency (cf. Paek & Cole, 2020; Tay, Meade, & Cao, 2015).

- *Researchers are encouraged to triangulate conclusions about the degree of model-data fit using multiple tests of fit (e.g., changes in LR, changes in R^2, information criteria, M_2).*

1.5.3.4 Item-Level and Person-Level Tests of Model-Data Fit

In addition to computing the omnibus measures of model-data fit, it is also recommended that researchers estimate the fit between the model and individual items and between the model and individual examinees.

Item-fit. There are a large number of statistics available to test model-data fit at the level of individuals items, denoted ***item fit.*** These statistics include: χ^2(Bock, 1972) and the Q1 variant offered by Yen (1981), $S-\chi^2$ (Orlando & Thissen, 2000), *Zh* (Drasgow, Levine, & Williams (1985), and G^2 (McKinley & Mills, 1985). Although there is no universally agreed upon statistic, several authors have recommended computing the Q1 statistic or the $S-\chi^2$ statistic (Mair, 2018; Orlando & Thissen, 2003; Paek & Cole, 2020). In addition, it is possible to visually compare the fit between estimated ICCs and empirical ICCs (see pp. 234-235 in Embretson & Reise, 2000 for additional information).

- *Researchers should supplement omnibus tests of model-data fit with tests of item-fit.*

- *Items flagged as problematic should either be removed from the test and the analyses repeated.*

Person-fit. As Embretson and Reise (2000) noted, "There are several dozen published and researched person-fit statistics…[but all of these indices] are based, in some way, on the consistency of an individual's item response pattern with some proposed model of valid item responding" (p. 238). A slightly tweaked interpretation of these ***person-fit*** statistics is that they are assessing the extent to which a person's item response pattern is inconsistent with the model that is being used to estimate their pattern of item responses. Person-fit statistics consider whether the proposed model (e.g., 1PL) does a good job representing each individual person's data. If there is significant mis-fit, it suggests that the IRT model is not working for that person. It is normal to have a few extreme scores on person-fit statistics when dealing with large samples. However, if a great many examinees have large person-fit statistics, then this provides additional evidence that the wrong IRT model is being fit to the data. When framed in this light, person-fit statistics may be thought of as providing an index of *appropriateness measurement* – whether the IRT measurement model is an appropriate representation of the individual's response pattern (de Ayala, 2009). One of the most effective indices of person-fit is Drasgow, Levine, and McLaughlin's (1987) l_Z statistic (see also Levine & Drasgow, 1983). *Zh* is the standardized version of l_Z. It has a conditional null distribution that is standard normal (i.e., mean of 0 and standard deviation of 1). Thus, scores on the *Zh* index may be compared against the *Z* values in a standard normal table to identify individuals with particularly unusual (i.e.,

problematic) response patterns. As de Ayala (2009) noted, "In general, a "good" l_Z is around 0.0. An l_Z that is negative reflects a relatively unlikely response vector (i.e., inconsistent responses), whereas a positive value indicates a comparatively more likely response vector than would be expected on the basis of the model (i.e., hyperconsistent responses)" (p. 143). Likewise, Paek and Cole (2020) suggested that individuals with *Zh* values greater than 3 "needs attention for a possibility of important aberrant response patterns" (p. 57).

- *Researchers should supplement omnibus tests of model-data fit with tests of person-fit.*
    - *Individuals whose item response pattern is identified as having poor-fit should be closely examined to determine the cause of misfit and removed if necessary, as would be seen with a group by construct interaction that would suggest differential item functioning.*

1.5.3.5 Estimating Latent Traits

Typically, maximum likelihood estimation (MLE) is used to generate the parameter estimates corresponding to the models described above (Baker & Kim, 2017; de Ayala, 2009; Embretson & Reise, 2000). After researchers have confirmed the fit between the data and their IRT model, they may proceed to estimate *person latent trait* scores for each of the examinees. Several estimation options exist including: MLE, maximum a posteriori (MAP), and expected a posteriori (EAP). de Ayala (2009) notes that "All three approaches for estimating a person's location (MLE, EAP, MAP) treat the item parameters' estimates as "known" and ignore their estimation error when estimating θ" (p.77). One limitation associated with using MLE is that it is not able to provide estimates of θ for examinees with scores of 0 or perfect scores. In contrast, both EAP and MAP are able to compute person trait estimates ($i.e., \hat{\theta}$), even when examines obtain these extreme scores. For a more thorough discussion of the different estimation methods, the reader is directed to de Ayala (2009) and Embretson and Reise (2000).

- *Researchers should estimate latent trait scores using MLE, EAP, or MAP.*
    - *When examinee response patterns include a pattern where all items were incorrectly answered or all items were correctly answered, one of the Bayesian estimators should be used: EAP or MAP.*

1.5.3.6 Item-Person Maps

Because item parameters and person latent trait scores are scaled using a common metric, it is possible to visually examine the joint distribution of examinees and items using *item-person maps* (de Ayala, 2009). Specifically, these plots map the item difficulty parameters and the estimates of the latent traits onto a common metric. This allows researchers to better understand the distribution of the latent trait and item difficulties. These maps, along with the item information curves, may be used to help guide decisions about the inclusion or omission of particular items from the test battery. For example, item-person maps help to identify portions of the latent trait distribution that could benefit from additional items. Figure 7 contains an item-person map based on a sample of N = 1000 individuals who completed 5 items from the Law School Admissions Test. By examining this figure, we see that all of the items had location

parameters greater than 0. However, the majority of respondents appear to have theta values less than 1. If the goal of this test was to estimate a broad range of theta levels, then researchers should consider adding additional items, especially items with difficulty values (i.e., location parameters) falling between +1 and -2 theta.

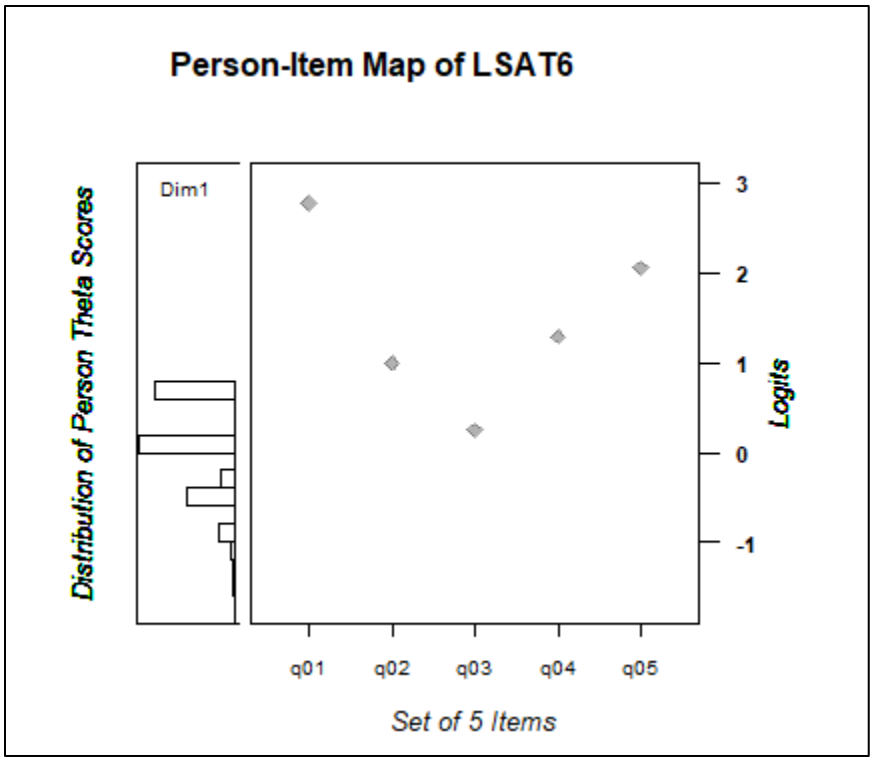


Figure 7. Example of Item-Person Map

As noted above:

- *When the target information function is uniform, researchers should strive to include roughly equal numbers of items from across the difficulty/theta continuum.*

- *When the target information function is peaked around a particular ability location (e.g., perhaps to increase the reliability of measurement around a specific cut score), researchers should strive to sample more items with difficulties matching the desired ability level.*

1.6 Item Bias and Test Bias

1.6.1. Definition of Bias

The term bias is typically interpreted by statisticians as implying the systematic over-or under-estimation of some focal parameter (e.g., means) as a function of group membership (e.g., majority vs. minority). Within the context of psychometrics, two forms of bias have been identified: structural bias and measurement bias.

1.6.1.1 Structural bias

When scores on a test (or item) have differential relationships with external variables across different groups, the test (or item) is said to be displaying structural bias (Embretson & Reise, 2000), which is also discussed under the rubric of differential validity (group differences in criterion-related validity correlations) or differential prediction (group differences in regression coefficients; predictive bias; Berry, 2015). Differential validity may be examined by computing correlations between the test and the criterion for each group and then testing whether those correlations differ from one another. Differential prediction may be examined using moderated multiple regression. In step 1, the criterion is regressed onto the test and group membership (either as a single dichotomous variable or as dummy/effect/contrast codes applied to a multi-category variable).

$$Y_i = \beta_0 + \beta_1(X_i) + \beta_2(G_i) + e_i \qquad (40),$$

where the βs denote unstandardized regression coefficients and Y_i, X_i, and G_i, represent person i's scores on the criterion, test, and group (0=White, 1 = Black) identifying variables, respectively.

In step 2, the cross-product between the test variable and the group membership variable(s) is added.

$$Y_i = \beta_0 + \beta_1(X_i) + \beta_2(G_i) + \beta_3(X_i G_i) + e_i \qquad (41).$$

A significant main effect for group membership indicates intercept differences whereas a significant cross-product(s) indicate slope differences. This form of bias may, or may not, be deemed problematic. For example, Berry's (2015) review of differential prediction of cognitive ability tests predicting performance revealed some limited evidence of prediction bias. However, this bias favored minority group members (e.g., Blacks) by overestimating their performance on the criterion and disfavored majority group members (e.g., Whites) by underestimating their performance on the criterion. In addition, Embretson and Reise (2000) noted that "…differential validity or lack of structural invariance may or may not be of concern depending on the context of test use. In many research contexts, the differential predictiveness of a measure is anticipated by a substantive theory, and may form the heart of a research program" (p. 250). However, absent a strong theory, items/tests displaying structural bias should be revised to eliminate this form of bias.

- *Researchers should test for differential validity (correlation coefficients) and differential prediction (regression coefficients) to determine if items/tests are displaying structural bias.*

    - *Problematic items/tests should be revised or removed, unless there is a compelling substance theory that justifies retaining the items.*

1.6.1.2 Measurement Bias

The second form of bias occurs when "…a test's internal relations (e.g., the covariances among item responses) differs across two or more groups of examinees" (Embretson & Reise, 2000, p. 250). More formally, *measurement bias* may be defined as:

> "When individuals who are identical on the construct measured by the test but who are from different subgroups have different probabilities of attaining the same observed score" (p. 438; Berry, 2015)

This form of bias is problematic because it suggests that the scale of measurement is not equivalent or invariant across groups. de Ayala (2009) summarized the process for determining whether an instrument suffers from measurement bias (see p. 324):

1. Examine the instrument at the item-level to identify potentially problematic items using tests of differential item functioning (DIF).
2. Items flagged as DIF should be reviewed by a panel of experts to ascertain the extent to which the DIF is relevant or irrelevant to latent construct. Such a review is referred to as the "logical evidence of bias."
3. Based on the panel's review, items may be revised or removed to eliminate measurement bias.

As noted above, the first step in evaluating items for measurement bias is to conduct a test of DIF. Under CTT, measurement bias is tested under the rubric of measurement invariance/measurement equivalence tests using confirmatory factor analyses (Vandenberg, 2002; Vandenberg & Lance, 2000). Under IRT, measurement bias is tested under the rubric of differential item/test functioning using a number of different test statistics (de Ayala, 2009; Embretson & Reise, 2000; Hambelton et al., 1991; Mair, 2018).

From an applied standpoint, consider measurement bias as a psychometric issue that indicates the test is actually measuring more than it is designed to test; for example, scores on a test might be a function of both the target construct (i.e., extroversion) and contaminated by construct irrelevant variance (i.e., race or gender). Think of structural bias as reflecting a statistical and theoretical issue that indicates the (purely measured) predictor construct has a different relationship with the criterion that is conditional on group membership. For example, I could have a test measuring depression that is psychometrically unbiased across a number of different demographic groups (e.g., race, gender, disability, veteran status, religion, etc.). However, I could find that this test has a stronger association with attempted suicide for certain groups (e.g., veterans with multiple combat tours). The test isn't biased for or against any group. However, the usefulness of the test differs across groups.

1.6.2. Testing for Measurement Bias: Choosing Between CTT and IRT Approaches

As noted above, frameworks exist for testing measurement bias using both the CTT and IRT psychometric models (Raju, Laffitte, & Bryne, 2002). Although some early research suggested these two frameworks may yield discrepant results, more recent research has suggested that discrepancies were likely engendered by the differential sequencing of the constraints/restrictions used to test for measurement bias rather than the reliance on CTT vs. IRT frameworks. Indeed, as Stark, Chernyshenko, and Drasgow (2006) concluded, when the pattern of constraints was held constant across the CTT and IRT frameworks, "…results indicated that CFA [CTT] and IRT were remarkably similar in their DIF detection accuracy" (p. 1303).

Given the potential for these two frameworks to yield similar results, this report will focus on tests of measurement bias under the IRT framework. Readers interested in additional resources on conducting tests of measurement invariance/measurement equivalence under the CTT model using CFA are directed to Stark et al., (2006), as well as earlier work by Vandenberg (2002), and Vandenberg and Lance (2000).

- *Researchers interested in testing for measurement bias under the CTT model are encouraged to follow the strategies presented in Vandenberg and Lance (2000), Vandenbeg (2002), and further clarified/refined in Stark et al. (2006).*

1.6.3. Important Considerations Using IRT to Test for Measurement Bias

There is a fundamental indeterminacy in the scale or metric of the latent constructs used in IRT models. This indeterminancy is of limited concern when researchers are analyzing data sampled from a single group (e.g., men). In such instances, the scale of the latent construct is typically (arbitrarily) set to have a mean of 0 and variance of 1. However, when transitioning to an analysis based on comparing data that are sampled from two or more groups (e.g., men and women) the scaling of the latent construct becomes a critical concern. As noted by Meade and Wright (2012):

> "For both IRT and confirmatory factor analytic [CTT] methods of invariance testing, there is an inherent indeterminancy associated with the metric of the latent variable…setting the metric is crucial for invariance analyses as there is an implicit assumption that the items selected as anchors are invariant across samples" (p. 1017).

1.6.3.1 Anchor Items

In order to conduct IRT-based DIF analyses, researchers must equate the metric of the latent construct across groups. This process of equating is accomplished by identifying an item (or ideally, a set of items) that are known to be free from DIF (i.e., items that are known to display measurement equivalence across groups). This set of DIF-free items, referred to as *anchor items,* is used by researchers to equate the metric of the latent constructs across groups. Because the metric of the latent construct is used to measure both people and items, the equating process typically involves a rescaling of item parameters prior to comparing the equivalence (or lack thereof) across groups.

In situations where solid anchor items are not known a priori, researchers must undertake an extra step to first identify and select items that will be used as the anchor items. Meade and Wright (2012) reviewed and compared several different strategies for identifying anchor items including:

1. All Other items As Anchors (AOAA)
   a. A baseline model is computed that constrains all item parameters to be equal across groups.
   b. Next, the equivalence constraints are relaxed for one item, while retaining all other items as anchors. The fit of this model is compared to the fit of the fully constrained model in step a. If the relaxed model shows better fit to the data, then the item is flagged as showing DIF. If the relaxed model fits the data as well as the fully constrained model, then it is said to be invariant across groups. This process is repeated separately for each of the K items.

2. ***Significance-based two-stage approach***
   a. The items judged to be invariant using the AOAA approach described above are used as anchor items.
   b. The remaining items (i.e., those flagged as DIF using AOAA) are retested for DIF using likelihood ratio tests (LRTs).

3. ***Fully iterative approach***
   a. This approach also builds off the AOAA approach. First, the AOAA approach is used and the item with the largest significant G^2 statistic is removed from the pool of items.
   b. The remaining items are then used in a new AOAA analysis. The item with the largest significant G^2 statistic is removed from the pool of items.
   d. The remaining items are then used again in a new AOAA analysis. This process continues until all items with significant G^2 statistics have been removed from the pool.
   e. The remaining items serve as anchor items and all previously discarded items are retested for DIF using LRTs.

4. ***maxA approach (two-stage approach based on item discrimination)***
   a. This two-stage approach conducts a preliminary DIF analysis using AOAA to identify a set of items free from DIF.
   b. From this set of DIF-free items, researchers select the items with the largest discrimination (a) parameter to serve as the anchor items and the remaining items are retested for DIF using LRTs.

5. ***minG^2 approach (two-stage approach based on items with smallest DIF statistic)***
   a. This two stage approach conducts a preliminary DIF analysis using AOAA to test for DIF (using the G^2 statistic).
   b. Items with the smallest G^2 statistic are selected to serve as anchor items and the remaining items are retested for DIF using LRTs.

6. ***minUIDS approach (two-stage effect-size based approach)***
   a. This two-stage approach conducts a preliminary DIF analysis using AOAA to test for DIF. The unsigned item difference in the sample (UIDS; Meade, 2010) is then computed

for all DIF-free items. This statistic is an effect size interpreted as the average absolute difference in expected scores across the sample of focal group respondents.

b. The non-significant items with the smallest effect sizes are then retained as anchor items and the remaining items are retested using LRTs.

Meade and Wright (2012) conducted a large Monte Carlo simulation to compare these strategies for identifying anchor items. They included three variations on each of the two-stage approaches. Specifically, using the decision heuristic for each of the two-stage approaches, they selected 1, 3, or 5 items to serve as the anchor items. Across their simulations, they found that larger sets of anchor items were associated with greater statistical power for detecting DIF. Across their initial simulation, they concluded that the maxA5 approach (i.e., two-stage approach that retains the 5 non-DIF items from the AOAA that have the largest discrimination parameter values) and the significance-based two-stage approach were most preferable. They then conducted a follow-up simulation to more closely scrutinize the performance of these two statistics. This led Meade and Wright (2012) to conclude, "…we unequivocally recommend" the maxA5 approach to identifying anchor items (p. 1028).

- *Researchers should use the maxA5 approach to identify anchor items when such items are not known on an a priori basis.*

1.6.3.2 Patterns of DIF

A large number of statistics are available to test items for DIF. Some statistics are able to detect ***uniform DIF,*** others, ***nonuniform DIF,*** and some both forms of DIF. When DIF is described as uniform, it implies that the graphs of the group ICCs do not crossover one another. Thus, uniform DIF is the IRT equivalent of an ordinal statistical interaction between group members and the latent construct in the prediction of item responses. Uniform DIF is engendered by similar discrimination parameters, but different difficulty parameters. As a result, the ICC for one group will be uniformly shifted up or down on the construct continuum relative to the ICC for the other group. In contrast, when DIF is described as nonuniform, it implies that the group ICCs do crossover one another. Nonuniform DIF is engendered by differences in discrimination parameters across groups.

1.6.3.3 Tests of DIF

A complete review of these statistics is outside the scope of this report However, Magis, Beland, Tuerlinckx, and De Boeck (2010) provided a nice overview of both IRT and non-IRT methods for detecting both uniform and non-uniform DIF, as presented in Table 8. In addition to reviewing these statistics, Magis et al. (2010) also provided a review of the difR package, which is an R package capable of implementing most of the statistics included in their review. A general summary of DIF tests was included in Table 1 of Magis et al. (2010), which is reproduced in Table 9. For additional information about the various tests of DIF, the reader is directed to the primary resources cited in Magis et al. (2010), as well as more general overviews in de Ayala (2009), Embretson and Reise (2000), Mair (2018), and Tay, Mead, and Cao (2015).

Table 9. Summary of DIF Tests

		Number of Groups	
Framework	**DIF Effect**	**2**	**>2**
Non-IRT	Uniform	Mantel-Haenszel* Standardization* SIBTEST Logistic regression*	Pairwise comparisons Generalized Mantel-Haenszel*
Non-IRT	Nonuniform	Logistic regression* Breslow-Day* NU.MH NU.SIBTEST	Pairwise comparisons
IRT	Uniform	LRT* Lord* Raju*	Pairwise comparisons Generalized Lord*
IRT	Nonuniform	LRT* Lord* Raju*	Pairwise comparisons Generalized Lord*

Note. NU.MH, modified Mantel-Haenszel for nonuniform DIF; NU.SIBTEST, modified SIBTEST for nonuniform DIF, LRT, likelihood ratio test. * Implemented in difR package (Version 2.2).

Reproduced from "A general framework and an R package for the detection of dichotomous differential item functioning", Magis, Beland, uerlinckx, and De Boeck (2010), Table 1, p. 849.

Although there are a number of different DIF statistics available, researchers have typically selected DIF statistics one of two ways. The first approach is to select a single DIF statistic that will serve as the tool used to judge items. One of the most commonly recommended statistics is the likelihood ratio test (cf. Meade & Wright, 2012; Tay et al., 2015). A second approach is to select a handful of tests (often combing both IRT and non-IRT based tests) and identifying problematic DIF items as those showing DIF across a majority of the tests. For example, Galic, Scherer, and LeBreton (2014) tested the Conditional Reasoning Test for Aggression for DIF by comparing samples from the US and Croatia. Each item was evaluated for DIF using 4 different tests: Lord's Chi-Square, Raju's Unsigned Area, Mantzel-Haenszel, and logistic regression. The authors noted that "it is common for different DIF criteria to lead to somewhat different conclusions (Borsboom, 2006), [thus] we decided to define as "true" DIFs those items for which the results of the four procedures converged" (p. 206).

- *Researchers are encouraged to test for DIF using either LRTs or a convergence based approach using multiple DIF tests drawn from both IRT and non-IRT traditions.*

1.6.3.4 Effect Sizes for DIF

Finally, researchers are encouraged to supplement statistical tests of DIF with estimates of DIF effect size. Meade (2010) provides an excellent review of the various effect size metrics that are available for use with tests of DIF. Specifically, he reviews six effect sizes applicable for use at the item-level and 9 effect sizes applicable for use at the test level-that is, ***differential test***

functioning examined using all items. Meade (2010) organized these statistics into a taxonomy of effect sizes based on the following factors (see Table 10):

- Is DIF allowed to cancel across items? Yes vs. No
- Is DIF allowed to cancel across respondents/theta? Yes vs. No
- What theta distribution is used? Sample distribution vs. assumed distribution

Table 10. Taxonomy of Item- and Scale-Level DIF Effect Sizes

	DF cancels across items?			
	Yes		**No**	
	DF cancels across respondents/thetas?		**DF cancels across respondents/thetas?**	
Theta Used	**Yes**	**No**	**Yes**	**No**
Sample	STDS ETSSD [a]	UETSDS Test D-Max Region of disadvantage Flowers et al.'s (1999) DTF	*SIDS* *ESSD* [a]	*UIDS* *Test D-Max* UTDS
Assumed Distribution	Stark's DTFR Stark's d_{DTF} [a]	UETDSN	*SIDN*	*UIDN* UDTFR

Note. Italics indicates an item-level index; normal font indicates a scale-level index. DF = differential functioning; STDS = signed test difference in the sample; UETSDS = unsigned expected test score difference in the sample; SIDS = signed item difference in sample; UIDS = unsigned item difference in sample; Test D-Max = maximum difference in expected test score for sample; ETSSD = expected test score standardized difference; DTF = differential tax functioning; ESSD = expected score standardized difference; UTDS=unsigned test difference in the sample; Stark's DTFR=Stark et al.'s (2004) DTFR; UETSDN = unsigned expected test score difference in normal distribution; SIDN = signed item difference in normal distribution; UIDN = unsigned item difference in normal distribution; Stark's d_{DTF} = Stark et al.'s (2004) d_{DTF}; UTDFR = unsigned DTFR.

[a] Indicates that index is standardized; other indices are in the metric of observed scores.

Reproduced from "A taxonomy of effect size measures for the differential functioning of items and scales", Meade (2010, Table 3, p. 734.

1.6.3.5 Recommended Steps in DIF Analysis

Based on a large review of the existing literature, Tay, Meade, and Cao (2015) summarized existing trends using IRT-based frameworks to test for measurement equivalence /differential item functioning. Following that review, the authors presented a set of general recommendations/steps researchers are encouraged to follow when testing for DIF. These general recommendations were then supplemented with specific recommendations for researchers relying on two specific IRT software programs-IRTPRO and Latent GOLD. Below I have summarized their general recommendations and I have also included specific recommendations (based on the work of Tay et al., 2015) that are also readily implemented using the R statistical package. Finally, as part of Appendix D, I have included an illustration of how to use an IRT framework to test for DIF using R.

1. *Researchers should begin by assessing model-data fit and unidimensionality.*
   a. Include a test of dimensionality (e.g., EFA, parallel analysis).
   b. Include an overall test of absolute model-data fit across all groups using the M2 statistic and/or the Root Mean Square Error pf Approximation (RMSEA) statistic computed using the M2 statistic (Maydeu-Olivares & Joe, 2006).
   c. Include a test of item-fit using the S-χ^2 (Orlando & Thissen, 2000) statistic and test the local dependence assuming using standardized LD-χ^2 (Chen & Thissen, 1997); tests should be conducted within each group.
   d. Include a test of relative model-data fit (e.g., 1PL vs. 2PL) using likelihood ratio tests or information criterion measures such as AIC and BIC
2. *Researchers should conduct an iterative DIF analysis*
   a. When anchor items are not known a priori, researchers should use empirical tests to identify DIF-free items to use as anchors (see Tay et al., 2015 for suggestions on identifying anchor items using IRTPRO and Latent GOLD; see previous section of the current report for general discussion of identifying anchor items).
3. *Estimate latent group mean differences*
   a. Researchers should constrain anchor items and freely estimate all DIF items prior to computing and interpreting latent group mean differences.
4. *Estimate effect sizes*
   a. Heuristic approach consists of 4 steps
      i. Estimate a fully constrained model; this model contains True mean level differences and Bias (Model TB)
      ii. Estimate a model with all DIF items freely estimated; this model contains True mean level differences only (Model T)
      iii. Examine the estimated latent trait difference between Model TB and Model T
      iv. Examine whether inferences about latent group mean differences vary across Model TB and Model T
   b. In addition, researches are encouraged to review Meade (2010) and compute additional item- and test-level effect sizes for DIF/DTF. Many of these effect sizes are readily implemented in the mirt package in R.

1.6.3.6 Conclusion

As is evident from this section, there exists a wide range of strategies that could be used to examine items and tests for evidence of differential item/test functioning using IRT-based models. As noted previously in the report, similar tools exist for examining measurement invariance/measurement equivalence using CTT-based models. As de Ayala (2009) suggested, evidence of DIF does not always imply measurement bias. There are some instances where DIF would be predicted by psychological theory. However, within the context of employment selection, items/tests displaying DIF are generally problematic and should be revised or eliminated from operational use.

2.0 REFERENCES

Air Force Examining Activities Overview-FY 2010-2011.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Prospect Heights, IL: Waveland Press.

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

Anastasi, A. (1968). *Psychological testing* (3rd ed.) Macmillan, Oxford.

Anderson, J. C., & Gerbing, D. W. (1991). Predicting the performance of measures in a confirmatory factor analysis with a pretest assessment of their substantive validities. *Journal of Applied Psychology, 76*(5), 732-740.

Andrich, D. (1979). A model for contingency tables having an ordered response classification. *Biometrics, 35*(2), 403-415.

Baker, F. B., & Kim, S. H. (2017). *The basics of item response theory using R*. Cham. Switzerland: Springer.

Benter, P. M., (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107(2)*, 238-246.

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.

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 74(3),* 478-494.

Binning, J. F., & LeBreton, J. M. (2009). Coherent conceptualization is useful for many things, and understanding validity is one of them. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 2(4),* 486-492.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29-51.

Brannick, M. T., Levine, E. L., & Morgeson, F. P. (2007). *Job and work analysis: Methods, research, and applications for human resource management* (2nd ed.) Sage Publications, Inc, Thousand Oaks, CA.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen and J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage Publications.

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(1),* 225-262.

Cascio, W. F., & Aguinis, H. (2019). *Applied psychology in talent management*. Thousand Oaks, CA: Sage.

Chen, W., & Thissen, D. (1997). Local independence for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22(3),* 265-289.

Cho, E. (2016). Making reliability reliable: A systematic approach to reliability coefficients. *Organizational Research Methods, 19*(*4),* 651-682.

Cho, E. & Kim, S. (2015). Cronbach's coefficient alpha: Sell known but poorly understood. *Organizational Research Methods, 18(2),* 207-230.

Colquitt, J. A., Sabey, T. B., Rodell, J. B., & Hill, E. T. (2019). Content validation guidelines: Evaluation criteria for definitional correspondence and definitional distinctiveness. *Journal of Applied Psychology, 104*(*10),* 1243–1265.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and application. *Journal of Applied Psychology, 78*(*1),* 98-104.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth.

Cronbach L. J. (1951) Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(*3),* 297-333.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. NY: Guilford Press.

DeShon, R. P. (1998). A cautionary note on measurement error corrections in structural equations models. *Psychological Methods, 3(4)*, 412-423.

DeShon, R. P. (2003). A generalizability theory perspective on measurement error corrections in validity generalization. In K. R. Murphy (Ed.), *Validity generalization: A critical review* (pp. 365-402). Mahwah, NJ: Lawrence Erlbaum.

DeSimone, J. (2015). New techniques for evaluating temporal consistency. *Organizational Research Methods, 18*(*1),* 133-152.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11,* 59-79.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38(1),* 67-86.

Duckworth, A. L., Peterson, C., Matthews, M., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. J*ournal of Personality and Social Psychology, 92*(*6),* 1087-1101. doi:http://dx.doi.org/10.1037/0022-3514.92.6.1087

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Fabrigar, L. R., Wegener, D. T., MacCullum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(*3),* 272-299.

Galic, Z., Scherer, K. T., & LeBreton, J. M. (2014). Examining the measurement equivalence of the conditional reasoning test for aggression across U.S. and Croatian samples. *Psychological Test and Assessment Modeling, 5*195-216.

Gatewood, R. D., Feild, H. S., & Barrick, M. R. (2016). *Human resource selection* (8th edition). Boston, MA: Cengage Learning.

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

Guilford, J. P., & Fruchter, B. (1973). *Fundamental statistics in psychology and education* (5th ed.). NY: McGraw-Hill.

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

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hayakawa, K. (2019). Corrected goodness-of-fit test in covariance structure analysis. *Psychological Methods, 24(3),* 371-389.

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods, 7(2),* 191–205. https://doi.org/10.1177/1094428104263675

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods, 1(1),* 104-121.

Hinkin, T. R., & Tracey, J. B. (1999). An analysis of variance approach to content validation. *Organizational Research Methods, 2*(*2),* 175-186.

Horn, J.L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179. https://doi.org/10.1007/BF02289447

James, L. R., LeBreton, J. M., Mitchell, T. R., Smith, D. R., DeSimone, J. A., Cookson, R., & Lee, H. J. (2013). Use of conditional reasoning to measure the power motive. In J. M.

Cortina, & R. S. Landis (Eds.), *Modern research methods for the study of behavior in organizations* (pp. 233-263). NY: Routledge/Taylor & Francis Group.

Johnson, J. & LeBreton, J. M. (2004). History and use of relative importance indices in organizational research. *Organizational Research Methods, 7*, 238-257.

Köhler, T., Cortina, J. M., Kurtessis, J. N., & Gölz, M. (2015). Are we correcting correctly?: Interdependence of reliabilities in meta-analysis. *Organizational Research Methods, 18(3),* 355-428.

Krasikova, D., LeBreton, J. M., & Tonidandel, S. (2011). Estimating the relative importance of variables in multiple regression models. In G. P. Hodgkinson & J.K. Ford (Eds.), *International review of industrial and organizational psychology, Vol 26* (pp. 119-141). Indianapolis, IN: Wiley.

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods, 9*(2), 202-220.

Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist, 41(11),* 1183-1192.

LeBreton, J. M., Burgess, J. R. D., Kaiser, R. B., Atchley, E. K. P., & James, L. R. (2003). The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar? *Organizational Research Methods, 6*, 80–128.

LeBreton, J. M., Hargis, M. B., Griepentrog, B., Oswald, F. L., & Ployhart, R. E. (2007). A multidimensional approach for evaluating variables in organizational research and practice. *Personnel Psychology, 60*, 475-498.

LeBreton, J. M., Ployhart, R. E., & Ladd, R. T. (2004). A Monte Carlo comparison of relative importance methodologies. *Organizational Research Methods, 7*, 258-282.

LeBreton J. M., & Senter J. L. (2008). Answers to twenty questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*, 815-852.

LeBreton, J. M, Tonidandel, S. (2008). Multivariate relative importance: Extending relative weight analysis to multivariate criterion spaces. *Journal of Applied Psychology, 93*, 329-345.

LeBreton, J. M., Tonidandel, S., & Krasikova, D. V. (2013). Residualized relative importance analysis: A technique for the comprehensive decomposition of variance in higher-order regression models. *Organizational Research Methods, 16*(3), 449-473.

Levine, M. V., & Drasgow, F. (1983). Appropriateness measurement: Validity studies and variable ability models. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computer adaptive testing* (pp. 109-131). NY: Academic Press.

Lievens, F., & Sackett, P. R. (2007). Situational judgment tests in high-stakes settings: Issues and strategies with generating alternate forms. *Journal of Applied Psychology, 92,* 1043-1055.

Loken, E., & Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. *British Journal of Mathematical and Statistical Psychology, 63,* 509-525.

Long, J. S. (1983). *Confirmatory factor analysis*. Newbury Park, CA: Sage.

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement, 14*(2), 117-138.

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

Lozano, L. M., Garcia-Cueto, E., & Muniz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology, 4*(*2),* 73-79.

Mair, P. (2018). *Modern Psychometrics with R*. NY: Springer.

Magis, D., Beland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavioral Research Methods, 42*(*3),* 847-862.

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika 47*(*2),* 149-174.

Maydeu-Olivares, A. & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*, 713. https://doi.org/10.1007/s11336-005-1295-9

McDonald, R. P. (2000). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement, 9*(*1),* 49–57. https://doi.org/10.1177/014662168500900105

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*(*3),* 412-433.

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology, 95*(*4),* 728-743.

Meade, A. W., & Wright, N. A. (2012). Solving the measurement invariance anchor item problem in item response theory. *Journal of Applied Psychology, 97*(*5),* 1016–1031. https://doi.org/10.1037/a0027934

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(*9),* 741-749.

Morgeson, F. P., & Campion, M. A. (1997). Social and cognitive sources of potential inaccuracy in job analysis. *Journal of Applied Psychology, 8*627-655.

Mulaik, S. A. (2009). *Foundations of factor analysis* (2nd ed.). NY: CRC Press/Taylor & Francis Group.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin, 105*(*3),* 430-445.

Murphy, K. R., & DeShon, R. (2000). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology, 53*, 873-900.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. Applied *Psychological Measurement, 16(2),* 159-176.

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

Nunnally, J. C., & Bernstein, R. H. (1994). *Psychometric theory*. NY: McGraw-Hill.

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology, 78*, 679-703.

Orlando, M., & Thissen, D. (2000). New item fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24(1),* 50-64.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X^2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement, 27(4),* 289-298.

Paek I., & Cole, K. (2020). *Using R for item response theory model applications*. NY: Routledge/Taylor & Francis Group.

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2016). Recommendations for creating better concept definitions in the organizational, behavioral, and social sciences. *Organizational Research Methods, 19(2),* 159-203.

Principles for the Validation and Use of Personnel Selection Procedures. (2018). Industrial and Organizational Psychology, 11(S1), 1-97. doi:10.1017/iop.2018.195

Putka, D. J., Le, H., McCloy, R. A., & Diaz, T. (2008). Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. *Journal of Applied Psychology, 93*, 959-981.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87(3),* 517-529.

Roth, P. L., Le, H., Oh, I., 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(5),* 802-828.

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology, 8*112-118.

Samejima, F. (2010). The general graded response model. In M. L. Nering, & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 77-107). NY: Routledge/Taylor & Francis Group.

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1(2),* 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*,262-274.

Schmidt, F. & Hunter, J. (2015). Meta-analysis of correlations corrected individually for artifacts. In Schmidt, F., & Hunter, J. *Methods of meta-analysis* (pp. 87-164). London: SAGE Publications, Ltd. doi: 10.4135/9781483398105

Schmidt, F. L., Hunter, J. E., & Urry, V. W. (1976). Statistical power in criterion-related validation studies. *Journal of Applied Psychology, 61(4),* 473-485.

Schmidt, F. L, Oh, I., & Le, H. (2006). Increasing the accuracy of corrections for range restriction: Implications for selection procedure validities and other research results. *Personnel Psychology, 59,* 281-305.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*(*2),* 93-105.

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

Smith, D. R., Hoffman, M. E., & LeBreton, J. M. (2020). Conditional Reasoning: An integrated approach to item analysis. *Organizational Research Methods, 23(1),* 124-153.

Shaffer, J. A., DeGeest, D., & Li, (2016). Tackling the problem of construct proliferation: A guide to assessing the discriminate validity of conceptually related constructs. *Organizational Research Methods, 19(1),* 80-110.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91(6),* 1292-1306.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston: Allyn & Bacon.

Tatsuoka, M. M., & Lohnes, P. R. (1988). *Multivariate analysis: Techniques for educational and psychological research* (2nd ed.). Macmillan Publishing Co, Inc.

Tay, L., Meade, A. W., & Cao, M. (2015). An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods, 18(1),* 3-6. https://doi.org/10.1177/1094428114553062

Tonidandel, S., & LeBreton, J. M. (2010). Determining the relative importance of predictors in logistic regression: An extension of relative weights analysis. *Organizational Research Methods, 13*, 767-781.

Tonidandel, S., & LeBreton, J. M. (2011). Relative importance analysis-A useful supplement to regression analyses. *Journal of Business and Psychology, 26*, 1-9.

Tonidandel, S., & LeBreton, J. M. (2015). RWA-Web: A free, comprehensive, web-based, and user-friendly tool for relative weight analyses. *Journal of Business and Psychology, 30(2),* 207-216.

Tonidandel, S., LeBreton, J. M.,, & Johnson, J. W. (2009). Statistical significance tests for relative weights. *Psychological Methods, 14,* 387-399.

Tonidandel, S., Williams, E. B., & LeBreton, J. M. (2015). Size matters…just not in the way that you think: Myths surrounding sample size requirements for statistical analyses. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends (Vol. 2): Doctrine, verity, and fable in the organizational and social sciences* (pp. 162-183). NY: Routledge.

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5*(2), 139-158.

Vandenberg, R. J., & Lance, C. E. (2000). Review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3(1),* 4-70.

Velicer, W. F., & Fava, J. L. (1998). Effects of variable and subject sampling on factor pattern recovery. *Psychological Methods, 3(2),* 231-251.

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81(5),* 557-574.

Wakita, T., Ueshima, N., & Noguchi, H. (2012). Psychological distance between categories in the Likert scale: Comparing different numbers of options. *Educational and Psychological Measurement, 72(4),* 533-546).

Widaman, K. F., & Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological Methods, 8*(*1),* 16-37.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30,* 187-213.

Yuan, Z., Morgeson, F. P., & LeBreton, J. M. (in press). Maybe not so independent after all: Exploring meta-analytic assumptions about the relationship between situational moderators and criterion reliability. *Personnel Psychology*.

LIST OF SYMBOLS, ABBREVIATIONS, AND ACRONYMS

1PL	1-parameter logistic
2PL	2-parameter logistic
3PL	3-parameter logistic
4PL	4-parameter logistic
A	ASVAB Administrative composite
AERA	American Educational Research Association
AF/A1PT	Air Staff, Air Force Testing Policy
AFEAO	Air Force Examining Activities Overview
AFECD	Air Force Enlisted Classification Directories
AFHRL	Air Force Human Resources Laboratory
AFOCD	Air Force Officer Classification Directories
AFS	Air Force Specialty
AFSC	Air Force Specialty Code
AFMAN	Air Force Manual
AFOQT	Air Force Officer Qualifying Test
AFPC//DP3SP	Air Force Personnel Center, Promotions, Evaluations, and Recognition branch
AFPC/DSYX	Air Force Personnel Center, Strategic Research and Analysis branch
AFPD	Air Force Policy Directive
AFRS	Air Force Recruiting Service
AF-WIN	Air Force Work Interest Navigator
AIC	Akaike information criterion
AOAA	All Other items As Anchors
ASVAB	Armed Services Vocational Aptitude Battery
ATC	Air Traffic Control
ATST	Air Traffic Scenarios Test
BIC	Bayesian information criterion
CFA	confirmatory factor analysis
CFM	career field manager
CM	competency model
CTT	classical test theory
DIF	differential item functioning
DoD	Department of Defense
E	ASVAB Electronics composite
EAP	expected a posteriori
EDPT	Electronic Data Processing Test
EFA	exploratory factor analysis
EPQT	Enlisted Pilot Qualifying Test
FY	Fiscal Year
G	ASVAB General composite
ICC	item characteristic curve
I/O	industrial/organizational
IRT	item response theory
KSA	knowledge, skills, and abilities

KSAO	knowledge, skills, abilities, and other characteristics
LF	maximum of the likelihood for the unrestricted or full model
LR	maximum of the likelihood function for the restricted model
LRT	likelihood ratio test
M	ASVAB Mechanical composite
MEPS	Military Entrance Processing Station
METS	Military Entrance Test Site
MAP	maximum a posteriori
MLE	maximum likelihood estimate
MTT	Multi-Tasking Test
OA	occupational analysis
PCA	principal component analysis
PCSM	Pilot Candidate Selection Method
RPA	Remotely Piloted Aircraft
RMSEA	Root Mean Square Error of Approximation
SDI	Self-Description Inventory
SEE	standard error of estimate
SME	subject matter expert
SRMR	standardized root mean-square residual
TAPAS	Tailored Adaptive Personality Assessment System
TBAS	Test of Basic Aviation Skills
UIDS	unsigned item difference in the sample
USAF	United States Air Force

APPENDIX A: Steps in Item Analysis and Test Evaluation Using CTT

Step 1 – Data management

- Open RStudio and load relevant packages
- Read data into R & verify integrity of data
    - Include relevant variables (e.g., demographics; subject ID; criteria; other predictors)
    - Format data: each row is a distinct examinee; each column is a distinct item or variable
- Identify columns with focal items

Step 2 – Item Analysis

- Item difficulty, Item discrimination and item-total correlations, Item validity

Step 3 – Scale Level Analysis

- Estimate reliability
- Exploratory and/or Confirmatory Factor Analysis; if the latter, be sure to use appropriate cutoffs for fit statistics (see Lance, Butts, and Michels, 2006).
- Modification Indices

Step 4 – Decide which items to retain as part of the test

- Using the results of item analyses in step 3, select items from the item bank that will address goals/purpose of the test (e.g., items that will maximize reliability vs. maximize validity vs. optimally discriminate at a particular point on the construct continuum).
- After removing items, iteratively re-compute relevant statistics (e.g., item-total correlations; convergent validity correlations).

Step 5 - Test for differential item/test functioning

- Use confirmatory factor analysis and tests of measurement equivalence/invariance to test for differential item functioning.
    - For detailed suggestions the reader is directed to Vandenberg & Lance (2000); Vandenberg (2000); Raju, Laffitte, & Byrne (2002); Stark, Chernyshenko, & Drasgow (2006)
- Remove problematic items from the item bank

Step 6 – Collect Additional Data & Accumulate Validity Evidence for Test (see Section II and III in report)

Step 7 – Replication/Cross-Validation

Step 8 – Develop norms, etc.

APPENDIX B: Steps in Item Analysis and Test Evaluation Using IRT

Step 1 – Data management

- Open RStudio and load relevant packages
- Read data into R & verify integrity of data
    - Include relevant variables (e.g., demographics; subject ID; criteria; other predictors)
    - Format data: each row is a distinct examinee; each column is a distinct item or variable
- Identify columns with focal items

Step 2 – Item Analysis: Test for Unidimensionality

- Parallel Analysis; modified parallel analysis; factor analysis; very simple structure analysis

Step 3 – Select IRT Model based on Fit Between Model and Data

- Run relevant IRT models
- Test-level evaluations of absolute fit: M2; SRMSR; etc.
- Test-level evaluations of relative model fit: -2LL test for nested models
- Item-level evaluations of fit: chi square test (with and without Bonferroni corrections)
- Person-level fit: lz
- Pairwise test of local independence: LD-chi square

Step 4 – Select & Use Model

- Calibrate the model – estimate parameters
- Generate person ability estimates (e.g., factor scores)
- Generate graphs of ICCs, TCC, IICs, TIC, and SEE

Step 7 – Decide which items to retain as part of the test

- Using IICs, select items from the item bank that will address the most difficult aspects of the target.
- After adding each item, iteratively re-compute the test information function.
- Continue adding items until the estimated test information function closely approximates the desired target information function.

Step 8 - Test for differential item/test functioning

- Use existing anchor items or identify anchor items using maxA5
- Test for differential item functioning using LRT
- Compute effect sizes for DIF/DTF
- Remove problematic items from the item bank

Step 9 – Collect Additional Data & Accumulate Validity Evidence for Test (see Section II and III in report)

Step 10 – Replication/Cross-Validation

Step 11 – Develop norms, etc.

APPENDIX C: Examples of CTT Item Analysis in R

```
##############################
###CTT Part 1
##############################
#Step 1 load  data
#Note: I tried analyzing the data John sent, but there were only n=74
#####  I duplicated the data to create a larger dataset, but encountered problems
##### when trying to use that data in the item analysis.
#####  Thus, I am going to illustrate the use of R for item/test analysis using
#####  cognitive data included in the psych package
getwd()
setwd("D:/Users/James/Dropbox/James Work Files/professional - consulting/PDRI/James' Chapter Drafts")

#Step 1a: Load relevant packages
install.packages("psych") #download the psych package to computer
library(psych) #load psych package into library of active resources

#Step 1b: load data from the psych package
data(package="psych")  #lists the data sets included in the psych package

#get information about the lsat6 dataframe
?lsat6

#read dataframes into r
lsat6=data.frame(lsat6)
lsat7=data.frame(lsat7)

#combine dataframes
df=cbind(lsat6,lsat7)

#add a subject ID number to data frame
df$id <- seq.int(nrow(df))

#The id variable now appears in column 11
names(df)

#This ID variable was created as an integer variable, but we want it to be a nominal variable
class(df$id)
```

```
#Overwrite the original integer version of id with correct nominal version, by converting it using
"as.factor" function;
df$id=as.factor(df$id); class(df$id)

#Change the names of the variables
names(df)[1] <- "q01"
names(df)[2] <- "q02"
names(df)[3] <- "q03"
names(df)[4] <- "q04"
names(df)[5] <- "q05"
names(df)[6] <- "q06"
names(df)[7] <- "q07"
names(df)[8] <- "q08"
names(df)[9] <- "q09"
names(df)[10] <- "q10"

head(df);tail(df) #print the first 6 rows and last 6 rows of the dataframe to verify things look okay
str(df) #request the structure of the df

#simulate criterion variable
df$y=rnorm(1000,m=50, sd=1)
df$y=df[,'y']+.5*df$q01+.5*df$q02+.5*df$q03+.5*df$q06+.5*df$q08
names(df)
head(df)
str(df)
cor(df[,c(12,1:10)])
#reorder columns in df
df=df[,c(11,12,1:10)]; names(df)
rm(lsat6,lsat7) #remove the original lsat dataframes from the global environment

write.table(df,file="df.csv",sep = ",", row.names=F) #write the dataframe to a csv file so results can be
compared using other software

# create a list containing variable namesAdded a couple of new keys that are based on fewer items
keys.list = list(
  all=c("q01","q02","q03","q04","q05","q06","q07","q08","q09","q10"),
  lsat6=c("q01","q02","q03","q04","q05"),
  lsat7=c("q06","q07","q08","q09","q10"))
#Open the next code file: Appendix C-CTT Final Part 2.R
```

```
##############################
###CTT Part 2
##############################
library(psych)

#############################################################################
#Step 1: Read in data/Data Managaement  (see Appendix D-IRT Final Part 1.R)
#############################################################################
library(psych)

### Create a list containing variable namesAdded a couple of new keys that are based on fewer items
keys.list = list(
  all=c("q01","q02","q03","q04","q05","q06","q07","q08","q09","q10"),
  lsat6=c("q01","q02","q03","q04","q05"),
  lsat7=c("q06","q07","q08","q09","q10"))

### Descriptive statistics for all variables in the dataframe
describe(df[,])

### Or just the columns with the items
names(df)
(item.descriptives=describe(df[,c(3:12)]))

### Use the keys to score the data and create scale scores
help(scoreItems) #provides information about the function scoreItems
help(make.keys)
keys=make.keys(df,keys.list)

### scoreItems will create composite scores, estimate item analysis statistics, & impute missing data
### scoreFast just scores the items with or without imputing missing data, but omits item-analysis
statistics
### scoreVeryFast only scores items -- no item analysis or imputation of missing
scores=scoreItems(keys,df,totals=TRUE)

### Basic summary of scores
scores

### List the elements stored in the scores object
names(scores)
```

```
### Merge scale scores back with original data file
# Note: cbind should only be used when you know the order of the files is identical
# If merging dataframes with different order of cases or different numbers of cases, use "merge" & select
the appropriate key variable
# help(merge)
data=cbind(df,scores$scores) #performs a column bind by adding scores to the right columns of the data
frame

#Verifying that scale scores were computed and added to the last 2 columns of data
names(data) #Verify that the last 3 columns now contain the scale scores
#Step 2: Item-Level Analyses
############################################################

### Item-Total Correlations: From this point foward will just focus on the set of lsat6 items
### Examine Structure of the elements in the scores data object
scores$item.cor

### From the above we see that column 1 contains correlations between the 10 items and the scale score
based on "all" (i.e., 10) items
### Column 2 contains correlations between the 10 items and the scale score based on the "lsat" scale score
(i.e. item 1-5); thus,
### only the first 5 correlations are relevant for the lsat6 item-total correlations; Column 3 contains
correlations for lsat7; thus, only
### the last 5 correlations are relevant for the lsat7 item-total correlations

scores$item.cor[c(1:5),2] #Item-total point-biserials; just pulling the relevant correlations from column 2
(lsat6 with q01 to q05)
scores$item.corrected[c(1:5),2] #Corrected item-total point-biserials for math items
(lsat6.biserials=biserial(data$lsat6, data[,keys.list$lsat6])) #Item-total biserials for the lsat6 items

###ITEM DISCRIMINATION INDEX
# Identify the individuals in the upper and lower 10th percentiles on lsat6 (not really meaningful with
only 5 items, but included for illusration)
quantile(data$lsat6,c(.10, .90))

# Create data frames containing only the examinees with verbal scores >= 90th percentile
lsat6.upper=data[which(data$lsat6>=5),]

# Create data frames containing only the examinees with verbal scores >= 90th percentile
lsat6.lower=data[which(data$lsat6<=2),]
```

```
# Compute the p-values within the upper and lower groups
lsat6.upper.mn=colMeans(lsat6.upper[,c(3:12)]) # Compute means for items (columns 3-12) for upper group
lsat6.lower.mn=colMeans(lsat6.lower[,c(3:12)]) # Compute means for items (columns 3-12) for lower group

#Estimate the item discrimination index (p_upper - p_lower) for each item
(lsat6.disc.index=lsat6.upper.mn-lsat6.lower.mn)

# Compute the ITEM-CRITERION CORRELATION (Biserials)
(lsat6.item.criterion=biserial(data$y, data[,keys.list$lsat6]))

###ASSEMBLE ITEM ANALYSIS SUMMARY
#Just pulling the statistics that are relevant to the math items
#Pulling some data from the item.descriptives dataframe; only need info for items in rows 1 - 5 (q01 to
q05)
item.descriptives
lsat6.summary=data.frame(item=keys.list$lsat6,
                         n=item.descriptives[1:5,2],
                         item.difficulty=item.descriptives[1:5,3], #Column 3 contains means/difficulties
                         item.variance=item.descriptives[1:5,4], #Column 4 contains variance
                         item.skew=item.descriptives[1:5,11], #Column 11 contains skew
                         item.kurtosis=item.descriptives[1:5,12], #Column 12 contains kurtosis
                         item.total.pbs=scores$item.cor[1:5,2], #Column 2 contains lsat6 correlations
                         item.total.pbs.corrected=scores$item.corrected[1:5,2], # Column 2 = contains
lsat6 correlations
                         item.total.biserials=lsat6.biserials,
                         item.disc.index=lsat6.disc.index[1:5],
                         item.criterion.biserial=lsat6.item.criterion)

### Review Summary of Item-Level Anaysis
print(lsat6.summary)
#Proceed to Scale Level Analyses...Appendix C-CTT Part 3.R
```

```
############################
###CTT Part 3
############################

###Part 3: Scale-Level Analysis
install.packages("GPArotation")
library(GPArotation)

#Reliability Estimates for each scale
scores$alpha #Cronbach's Coefficient Alpha
scores$G6 #Guttman's Lambda 6

#splitHalf function in psych package provides a number of additional estimates of reliability
help(splitHalf)
lsat6.reliability=splitHalf(data[,keys.list$lsat6]); lsat6.reliability

#Estimate omega using function in psych package
#Assuming there is a single dominant factor; warnings b/c I copied & pasted small dataset to create large
data set
lsat6.omega=omega(data[,keys.list$lsat6],nfactors=1)
lsat6.omega #Look at the estimate of omega total if specifying 1 factor

#Correlations with external variables included in dataframe (e.g., Class Rank)
names(data)
cor(data[,c("y","lsat6")])

#Exploratory Factor Analysis
lsat6.parallel <- fa.parallel(data[,keys.list$lsat6], fm = 'pa')

# Examine the output and the graph - probably 2-5 factors
help(fa) #Details about the fa function

# Estimating principal axis factor analysis with a single factor
lsat6.efa1=fa(data[,keys.list$lsat6], nfactors=1,SMC=T,rotate="oblimin",fm="pa", n.iter=1000)

# Pattern of factor loadings
lsat6.efa1

#Confirmatory Factor Analysis
#For tests that already have some prior validity & psychometric evidence, one could use CFA rather than EFA
```

```
install.packages("lavaan")
library(lavaan)

#Specified an arbitary 3 factor model using 9 items from previous EFA
#This is for illustrative purposes of the R code
#It is not appropriate to run CFA on same data as EFA
#Need to use EFA to refine scale, collect new data, then run CFA on new data

### Factors are allowed to correlate (could constrain correlations to zero using, F1 ~~ 0*F2
cfa.2f= '
F1_lsat6 =~ q01+q02+q03+q04+q05
F2_lsat7 =~ q06+q07+q08+q09+q10
F1_lsat6 ~~ F2_lsat7
'

### Estimate CFA and include constraint of factor variances to unity
fit.cfa.2f=sem(cfa.2f, data[,keys.list$all],std.lv=T); fit.cfa.2f

### Review model parameters and fit statisics
summary(fit.cfa.2f,fit.measures=T)

### Review modification indices
modindices(fit.cfa.2f)

### OTHER CONSIDERATIONS: Test Item Types
### Parallel Items within each of the three factors

### To constrain parameters, simply multiple them by the same constant/constraint

cfa.2f.par= '
F1_lsat6 =~ a*q01+a*q02+a*q03+a*q04+a*q05
F2_lsat7 =~ b*q06+b*q07+b*q08+b*q09+b*q10
F1_lsat6 ~~ F2_lsat7
q01~~c*q01
q02~~c*q02
q03~~c*q03
q04~~c*q04
q05~~c*q05
q06~~d*q06
q07~~d*q07
```

```
q08~~d*q08
q09~~d*q09
q10~~d*q10
'

fit.cfa.2f.par=cfa(cfa.2f.par, data[,keys.list$all],std.lv=T); fit.cfa.2f.par
summary(fit.cfa.2f.par,fit.measures=T)
### Tau Equivalent Items within each of the three factors
cfa.2f.tau= '
F1_lsat6 =~ a*q01+a*q02+a*q03+a*q04+a*q05
F2_lsat7 =~ b*q06+b*q07+b*q08+b*q09+b*q10
F1_lsat6 ~~ F2_lsat7'

fit.cfa.2f.tau=cfa(cfa.2f.tau, data[,keys.list$all],std.lv=T); fit.cfa.2f.tau
summary(fit.cfa.2f.tau,fit.measures=T)

#Compare Fit of Parallel Items vs. Tau Equivalent Items
anova(fit.cfa.2f.tau,fit.cfa.2f.par)

#Compare Fit of Tau Equivalent to original (Congeneric)
anova(fit.cfa.2f,fit.cfa.2f.tau)

#Tests of measurement equivalence/invariance are easily implmented using the lavaan package
#Excellent tutorial is available at: http://lavaan.ugent.be/tutorial/tutorial.pdf
```

APPENDIX D: Examples of IRT Item Analysis in R

```
#############################
###IRT Part 1
#############################

#Step 1 load  data
#Note: I tried analyzing the data John sent, but there were only n=74
#####  I duplicated the data to create a larger dataset, but encountered problems
##### when trying to use that data in the item analysis.
#####  Thus, I am going to illustrate the use of R for item/test analysis using
#####  cognitive data included in the psych package
getwd()
setwd("D:/Users/James/Dropbox/James Work Files/professional - consulting/PDRI/James' Chapter Drafts")

#Step 1a: Load relevant packages
install.packages("psych") #download the psych package to computer
library(psych) #load psych package into library of active resources

#Step 1b: load data from the psych package
data(package="psych")  #lists the data sets included in the psych package

#get information about the lsat6 dataframe
?lsat6

#read dataframes into r
lsat6=data.frame(lsat6)
lsat7=data.frame(lsat7)

#combine dataframes
df=cbind(lsat6,lsat7)

#add a subject ID number to data frame
df$id <- seq.int(nrow(df))

#The id variable now appears in column 11
names(df)

#This ID variable was created as an integer variable, but we want it to be a nominal variable
class(df$id)
```

```
#Overwrite the original integer version of id with correct nominal version, by converting it using
"as.factor" function;
df$id=as.factor(df$id); class(df$id)

#Change the names of the variables
names(df)[1] <- "q01"
names(df)[2] <- "q02"
names(df)[3] <- "q03"
names(df)[4] <- "q04"
names(df)[5] <- "q05"
names(df)[6] <- "q06"
names(df)[7] <- "q07"
names(df)[8] <- "q08"
names(df)[9] <- "q09"
names(df)[10] <- "q10"

head(df);tail(df) #print the first 6 rows and last 6 rows of the dataframe to verify things look okay
str(df) #request the structure of the df

#simulate criterion variable
df$y=rnorm(1000,m=50, sd=1)
df$y=df[,'y']+.5*df$q01+.5*df$q02+.5*df$q03+.5*df$q06+.5*df$q08
names(df)
head(df)
str(df)
cor(df[,c(12,1:10)])
#reorder columns in df
df=df[,c(11,12,1:10)]; names(df)
rm(lsat6,lsat7) #remove the original lsat dataframes from the global environment

write.table(df,file="df.csv",sep = ",", row.names=F) #write the dataframe to a csv file so results can be
compared using other software

# create a list containing variable namesAdded a couple of new keys that are based on fewer items
keys.list = list(
  all=c("q01","q02","q03","q04","q05","q06","q07","q08","q09","q10"),
  lsat6=c("q01","q02","q03","q04","q05"),
  lsat7=c("q06","q07","q08","q09","q10"))
```

```
#Open the next code file: Appendix D-IRT Final Part 2.R
##############################
###IRT Part 2
##############################

########## Item Response Theory: Item Analysis and Test Evaluation
###Part 2: Item Analyses
library(psych)
install.packages("GPArotation")
library(GPArotation)

### ITEM ANALYSIS: UNIDIMENSIONAL IRT MODELS

#################################################################
#Step 1: Read in data (see Appendix D-IRT Final Part 1.R)
#################################################################


#################################################################
#Step 2: Test for Unidimensionality
#################################################################

###Traditional parallel analysis
all.parallel <- fa.parallel(df[,keys.list$all], fm = 'pa')
lsat6.parallel <- fa.parallel(df[,keys.list$lsat6], fm = 'pa')
lsat7.parallel <- fa.parallel(df[,keys.list$lsat7], fm = 'pa')

###Modified parallel anlaysis available as part of the ltm package
install.packages("ltm")
library(ltm)

###Note: Modified PA requires first estimating an IRT model that can be used as comparison in simulation
all.ltm= ltm(df[,keys.list$all]~z1)
all.ltm.2fac= ltm(df[,keys.list$all]~z1+z2)
anova(all.ltm,all.ltm.2fac)
unidimTest(all.ltm) #This will take a while to run

lsat6.ltm = ltm(df[,keys.list$lsat6]~z1)
lsat6.ltm.2fac = ltm(df[,keys.list$lsat6]~z1+z2)
anova(lsat6.ltm,lsat6.ltm.2fac)
```

```
unidimTest(lsat6.ltm) #This will take a while to run

# Examine the output and the graph - looks like a single dominant factor within each of lsat6 and lsat7
# Combined might be multidimensional, but for our purposes we will assume a single dominant factor
underlies these 10 items

#################################################################
#Step 3: Model-Data Fit
#################################################################
# Going to run models in mirt package
install.packages("mirt")
library(mirt)
mirtCluster(4) #speeds things up

# Rasch Model using mirt (discrimination constrained to unity; difficulties freely estimated)
lsat6.rasch=mirt(df[,keys.list$lsat6],model=1,itemtype="Rasch",SE=T)
lsat6.rasch
coef(lsat6.rasch,IRTpars=T,simplify=T)
coef(lsat6.rasch,IRTpars=T)

#1PL using mirt (discrimination estimated and constraiend across items; difficulties freely estimated)
spec<-'all = 1-5
CONSTRAIN=(1-5,a1)' #estimating, but then constraining the slope across all items
lsat6.1pl<-mirt(df[,keys.list$lsat6], model=spec, itemtype="2PL", SE=T)
lsat6.1pl

# 2PL using mirt
lsat6.2pl=mirt(df[,keys.list$lsat6],model=1,itemtype="2PL",SE=T)
lsat6.2pl

# 3PL using mirt
lsat6.3pl=mirt(df[,keys.list$lsat6],model=1,itemtype="3PL",SE=T)
lsat6.3pl

# Examine Test-Level Model-Data Fit using -2LL Test, M2, RMSR, BIC, AIC, etc.

anova(lsat6.1pl,lsat6.2pl)
anova(lsat6.2pl,lsat6.3pl)

M2(lsat6.1pl)
```

```
M2(lsat6.2pl)
M2(lsat6.3pl)

# Examine Item-Level Evaluations of Fit
itemfit(lsat6.1pl)
itemfit(lsat6.2pl)
itemfit(lsat6.3pl)

# Examine Person-Level Fit
lsat6.1pl.pfit=personfit(lsat6.1pl)
lsat6.1pl.pfit=cbind(1:nrow(df),df,lsat6.1pl.pfit)
names(lsat6.1pl.pfit)

lsat6.2pl.pfit=personfit(lsat6.2pl)
lsat6.2pl.pfit=cbind(1:nrow(df),df,lsat6.2pl.pfit)
names(lsat6.2pl.pfit)

lsat6.3pl.pfit=personfit(lsat6.3pl)
lsat6.3pl.pfit=cbind(1:nrow(df),df,lsat6.3pl.pfit)
names(lsat6.3pl.pfit)

# Now sort the new person-fit data frame by Zh and examine dataframe for extreme scores
head(lsat6.1pl.pfit[order(lsat6.1pl.pfit[,18] ),]) #Zh is is column 18 of new dataframe b/c additional fit
stats are computed for 1PL
tail(lsat6.1pl.pfit[order(lsat6.1pl.pfit[,18] ),])

head(lsat6.2pl.pfit[order(lsat6.2pl.pfit[,14] ),]) #Zh is in column 14 for 2PL
tail(lsat6.2pl.pfit[order(lsat6.2pl.pfit[,14] ),])

head(lsat6.3pl.pfit[order(lsat6.3pl.pfit[,14] ),]) #Zh is in column 14 for 3PL
tail(lsat6.3pl.pfit[order(lsat6.3pl.pfit[,14] ),])

# Test of Pairwise Local Independence
# Compute LD-X2 using model residuals (Chen & Thissen, 1997)
lsat6.1pl.res=residuals(lsat6.1pl, type ="LD")
lsat6.2pl.res=residuals(lsat6.2pl, type ="LD")
lsat6.3pl.res=residuals(lsat6.3pl, type ="LD")

# Using residuals, compute LD-X2 (see the lower diagnoal of these matrices)
# Item pairs with values > 10 may
```

```
(abs((lsat6.1pl.res)-1)/sqrt(2))
(abs((lsat6.2pl.res)-1)/sqrt(2))
(abs((lsat6.3pl.res)-1)/sqrt(2))

# Q3 may also be estimated
residuals(lsat6.1pl, type ="Q3")
residuals(lsat6.2pl, type ="Q3")
residuals(lsat6.3pl, type ="Q3")

################################################################
#Step 4: Select & Use Model
################################################################
# All models had good fit to the data
# I will retain 2PL for rest of this section

# Calibrate the 2PL
lsat6.2pl=mirt(df[,keys.list$lsat6],model=1,itemtype="2PL",SE=T)

# Examine overall results
lsat6.2pl

# Examine parameter estimates (some different options for viewing results)
# Note: a = discrimination; b= difficulty; g = guessing (lower asymptote); u = upper asymptote
#       for the 2PL, a & b are estimated, g is set to 0 and u is set to 1
coef(lsat6.2pl, IRTpars=T, simplify=T)
coef(lsat6.2pl, IRTpars=T)
coef(lsat6.2pl, IRTpars=T, printSE=T)

# Estimate latent trait scores
lsat6.2pl.scores=fscores(lsat6.2pl, method = "EAP", full.scores=T, full.scores.SE=T)
head(lsat6.2pl.scores); tail(lsat6.2pl.scores)

# Merge factor scores with original data
# cbind function will do the trick, assuming that the order of the original df has not be changed since the
2pl was estimated
df.final=cbind(df,lsat6.2pl.scores)
names(df.final)

# Generate plots of ICCs, TCC, IICs, TIC, SEE
# Plot a single ICC for item #4
```

```
itemplot(lsat6.2pl,4)

# Generate separate ICC plots for each item in lsat6
for (i in 1:length(keys.list$lsat6))
  {p=itemplot(lsat6.2pl,i)
print(p)
  }
# Generate a single plot of the ICCs for the items in lsat6
plot(lsat6.2pl,type="trace")

# Generate the Test Characteristic Curve
plot(lsat6.2pl)

# Generate a single Item Information Curve for item #4
itemplot(lsat6.2pl,4,type="info")

# Generate separate IIC plots for each item in lsat6
for (i in 1:length(keys.list$lsat6))
{p=itemplot(lsat6.2pl,i,type="info")
print(p)
}

# Generate Test Information Curve
plot(lsat6.2pl,type="info")

# Determine maximum information available by extracting info for theta=b)
coef(lsat6.2pl,IRTpars=T, simplify=T)
iteminfo(extract.item(lsat6.2pl,1), -3.361)
iteminfo(extract.item(lsat6.2pl,2), -1.370)
iteminfo(extract.item(lsat6.2pl,3), -0.280)
iteminfo(extract.item(lsat6.2pl,4), -1.866)
iteminfo(extract.item(lsat6.2pl,5), -3.123)

# Generate Person-Item Map
install.packages("WrightMap")
library(WrightMap)
names(df.final)
thetas.2pl <- df.final$F1 # create a new object containing estimates of theta
difficulties.2pl <- coef(lsat6.2pl,simplify=T)$items[1:5,2] # create a new object containing item
difficulties
```

```
wrightMap(thetas.2pl, difficulties.2pl,
          main.title = "Person-Item Map of LSAT6",
          axis.persons = "Distribution of Person Theta Scores",
          axis.items = "Set of 5 Items",
          show.thr.lab = F,
          show.thr.sys = F,
          item.side=itemModern,
          person.side=personHist)

wrightMap(thetas.2pl, difficulties.2pl,
          main.title = "Person-Item Map of LSAT6",
          axis.persons = "Distribution of Person Theta Scores",
          axis.items = "Set of 5 Items",
          show.thr.lab = F,
          show.thr.sys = F,
          item.side=itemModern,
          person.side=personDens)
```

```
##############################
###IRT Part 3
##############################

########## Item Response Theory: Item Analysis and Test Evaluation
###Part 3: ME/I DIF
library(psych)
library(GPArotation)

### MEASUREMENT EQUIVALENCE/INVARAIANCE; DIFFERENTIAL ITEM FUNCTIONING

#################################################################
#Step 1: Read in data (see Appendix D-IRT Final Part 1.R)
#################################################################
### I will use the data from the previous IRT appendices
### I will append a group variable to this data set
df$sex = rep(seq(0,1),500)

### Create separate dataframes for males and females
df.m = df[sex==0,]
df.f = df[sex==1,]
#################################################################
#Step 2: Test for Unidimensionality
#################################################################
# Will focus on the combined set of items (lsat6 + lsat7)

###Traditional parallel analysis
all.parallel <- fa.parallel(df[,keys.list$all], fm = 'pa')

###Modified parallel anlaysis available as part of the ltm package
install.packages("ltm")
library(ltm)

###Note: Modified PA requires first estimating an IRT model that can be used as comparison in simulation
all.ltm = ltm(df[,keys.list$all]~z1)
all.ltm.2fac = ltm(df[,keys.list$all]~z1+z2)
anova(all.ltm,all.ltm.2fac)
unidimTest(lsat6.ltm) #This will take a while to run

detach("package:ltm", unload = TRUE)
```

```
### The modified parallel tests suggests unidimensionality

################################################################
#Step 3: Model-Data Fit
################################################################

library(mirt)
mirtCluster(6) #speeds things up
# Assuming that the 2PL model is the appropriate model

### Overall model fit across groups (using M2)

### 2PL using mirt
all.2pl=mirt(df[,keys.list$all],model=1,itemtype="2PL",SE=T)
all.2pl

M2(all.2pl)

### Model fit within group S-X2
all.2pl.m = mirt(df.m[,keys.list$all],model=1,itemtype="2PL", SE=T)
all.2pl.f = mirt(df.f[,keys.list$all],model=1,itemtype="2PL", SE=T)

itemfit(all.2pl.m)
itemfit(all.2pl.f)

# Test of Pairwise Local Independence
# Compute LD-X2 using model residuals (Chen & Thissen, 1997)
all6.2pl.res.all=residuals(all.2pl, type ="LD")
all.2pl.res.m=residuals(all.2pl.m, type ="LD")
all.2pl.res.f=residuals(all.2pl.f, type ="LD")

# Using the above residuals residuals, compute LD-X2 (see the lower diagnoal of these matrices)
# Item pairs with values > 10 may
(abs((all.2pl.res.all)-1)/sqrt(2))
(abs((all.2pl.res.m)-1)/sqrt(2))
(abs((all.2pl.res.f)-1)/sqrt(2))

# Q3 may also be estimated
residuals(lsat6.2pl, type ="Q3")
```

```
residuals(lsat6.2pl.m, type ="Q3")
residuals(lsat6.2pl.f, type ="Q3")

################################################################
#Step 4: ITERATIVE DIF ANALYSIS
################################################################
### Assume: 2PL model is appropriate
### Assume: No information is avaialble on anchor items
### Plan: use Mead and Wright's (2012) suggestion to us the maxA5 approahc to get anchors

#################Freely Estimated Parameters###############
help(multipleGroup)
model.free <- multipleGroup(df[,keys.list$all], 1, as.factor(sex))
coef(model.free, simplify = TRUE) # for the manuscript

###Baseline Model
################## Baseline Model ############################
model.constrained <- multipleGroup(df[,keys.list$all], 1, as.factor(sex),
                                   invariance = c(colnames(df[,keys.list$all]), 'free_means', 'free_var'))
(constrained.parameters <- coef(model.constrained,simplify = TRUE)[[1]][[1]])

###First Round of LRTs
#################### First round of DIF analyses - All Others As Anchors #################
dif.drop <- DIF(model.constrained, c('a1','d'), scheme = 'drop', seq_stat = .05)

#########################################################################################
### RUN THE FOLLOWING FUNCTION TO FACILIATE ORGANIZING THE OUTPUT
get.dif.items <- function(f.data,p.val=.05,parms){
  r.warnings = ""
  keep.vars <- c("X2", "df", "p") # just keep these variables
  f.data <- f.data[keep.vars]
  f.data$p = round(f.data$p,3)
  if(missing(f.data)) return('Missing model output out.list')
  f.data$sig <- ifelse(f.data$p < p.val,'dif','no_dif')
  if(!missing(parms)){
    if(nrow(f.data) == nrow(parms)){
      f.data <- cbind(f.data,parms)
    }else{
      r.warnings = "There number of item parameters doesn't match the number of items "
      r.warnings = paste(r.warnings,"given to get.dif.items. Item parameters omitted.")
```

```
    }
  }
  dif.items <- subset(f.data, sig == 'dif')
  no.dif.items <- subset(f.data, sig == 'no_dif')
  if(!missing(parms) && nrow(f.data) == nrow(parms)){
    if(nrow(no.dif.items)>1){
      no.dif.items <- no.dif.items[order(-no.dif.items$a1),]
    }
  }
  r.list <- list(dif_items = dif.items, no_dif = no.dif.items, warnings = r.warnings)
  return(r.list)
}
##########################################################################################

## The above function let's us run the next line of code
get.dif.items(f.data=dif.drop,p.val=.05,parms=constrained.parameters)

###Specify a New Baseline Model using Anchor Items
###We will use the A5 method from Meade and Wright (2012) in which we will choose five anchor items with
###the largest A parameters. Note that the get.dif.items function will sort non-dif items by the A
parameter if
###supplied.
# q06
# q01
# q02
# q07
# q08

itemnames <- colnames(df[,keys.list$all])
anc.items.names <- itemnames[c(6,1,2,7,8)] #selected 5 non-dif items with largest a1 parameter
test.items <- c(3,4,5,9,10)
model_anchor <- multipleGroup(df[,keys.list$all], model = 1, group = as.factor(sex),
                              invariance = c(anc.items.names, 'free_means', 'free_var'))
(anchor.parms <-coef(model_anchor,simplify = TRUE)[[1]][[1]])

###Run the Final Invariance Tests
(dif.anchor <- DIF(model_anchor, c('a1','b'), items2test = test.items, plotdif = TRUE))

#################### Final round of DIF analyses #################
## use the optional function to table the output
```

```
get.dif.items(f.data=dif.anchor,p.val=.05)


###Step 7: Compute Effect Sizes
#The last step is to compute effect size estimates, as described in Meade (2010).
#Test-Level Effect Sizes
empirical_ES(model_anchor, DIF=FALSE,ref.group=1) # test-level effect sizes

empirical_ES(model_anchor,ref.group=2) # item-level effect sizes

empirical_ES(model_anchor, ref.group=2,DIF=FALSE, plot=TRUE) # expected test score plot

empirical_ES(model_anchor, ref.group=2,plot=TRUE)

itemplot(model_anchor, 9) # Plot item 9
```