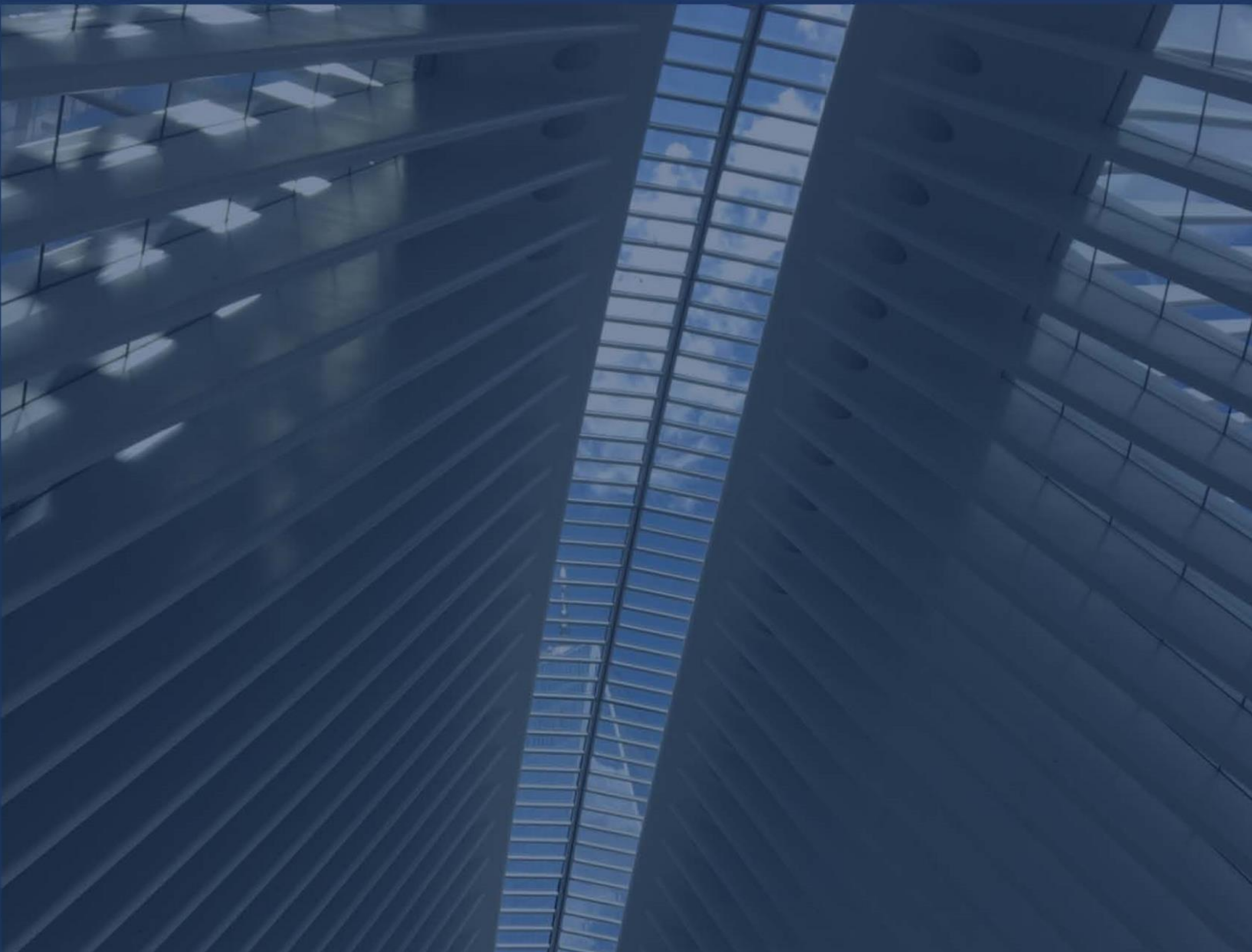
Guidelines for Technology-Based Assessment

International Test Commission and Association of Test Publishers



Published by the

Association of Test Publishers

601 Pennsylvania Ave., N.W., Suite 900

Washington D.C. 20004

and the

International Test Commission

77-79 The Courtyard

Hemel Hempstead

England

HP1 1LF

ISBN 979-8-88862-517-0

Cover photo © Stephen G. Sireci, 2023

Suggested Citation:

International Test Commission and Association of Test Publishers (2022). Guidelines for technology-based assessment.

https://www.intestcom.org/page/28 and https://www.testpublishers.org/white-papers

CONTENTS

PREFACE v

ACKNOWLEDGMENTS vi

PART I. INTRODUCTION AND BACKGROUND 1

Purpose of the *Guidelines* 1

Rationale and Salient Issues 1

Scope of the Guidelines 2

Audience 2

Foundational Documents and References 5

PART II. FOUNDATIONAL CONCEPTS AND CONSIDERATIONS 7

Validity and Fairness 7

Construct Underrepresentation/Construct-Irrelevant Variance 7

Reliability and Measurement Precision 8

Testing Contexts: High-Stakes, Low-Stakes 9

PART III. GUIDELINES FOR TECHNOLOGY-BASED ASSESSMENT 10

1. TEST DEVELOPMENT 10

Background 10

Planning for Technology-Based Assessments 10

Technology-Enhanced Items (TEIs) 11

Gamification, Game-Based Assessment, and Virtual Performance Assessments 13

Universal Test Design 14

Integrating Assessments with Instruction 15

Item Authoring 16

Guidelines for Test Development 17

Guidelines for Planning a Technology-Based Assessment 17

Guidelines for Technology-Enhanced Items 19

Guidelines for Development of Game-Based and Virtual Performance Assessment 21

Guidelines for Universal Test Design in Technology-Based Assessment 24

Guidelines for Developing Technology-Based Assessments Integrated with Instruction 24

Guidelines for Item Authoring and Review 26

2. TEST DESIGN AND ASSEMBLY.......... 28

Background.......... 28

Linear Test Design.......... 28

Adaptive Test Design.......... 29

Test Assembly.......... 30

Guidelines for Test Design and Assembly.......... 30

3. TEST DELIVERY ENVIRONMENTS.......... 34

Background.......... 34

Web-Based Delivery.......... 34

Offline, Local, and Mobile Delivery.......... 35

Locked-Down Browsers.......... 36

Interoperability.......... 37

Test Disruptions.......... 37

Guidelines for Test Delivery Environments.......... 38

Guidelines for Web-Based Delivery.......... 38

Guidelines for Offline, Local, and Mobile Delivery.......... 41

Guidelines for Locked-Down Browsers.......... 44

Guidelines for Interoperability.......... 48

Guidelines Addressing Test Disruptions.......... 51

4. SCORING.......... 55

Background.......... 55

Automated Scoring of Selected Responses.......... 55

Automated Scoring of Constructed Responses.......... 56

Technology-Assisted Human Scoring.......... 58

Scoring in the Event of Technology Disruption and Incomplete Assessments.......... 59

Using Item Response Time in Scoring.......... 59

Guidelines for Scoring Technology-Based Assessments.......... 60

Guidelines for Automated Scoring of Selected-Response Items.......... 60

Guidelines for Automated Scoring of Constructed-Response Items.......... 62

Guidelines for Technology-Assisted Human Scoring.......... 65

Guidelines for Scoring in the Event of Technology Disruption and Incomplete Testing Sessions.......... 66

Guidelines for Using Item Response Time in Scoring.......... 67

5. DIGITALLY BASED RESULTS REPORTING.......... 69

Background....69
Maintaining Confidentiality of Score Reporting....69
Guidelines for Digitally Based Results Reporting....70
Guidelines for Results Reporting....70
Guidelines for Quality Control in Score Reporting....71
Guidelines for Maintaining Confidentiality of Score Reporting....72
6. DATA MANAGEMENT....74
Background....74
Data Governance....74
Data Maintenance, Integrity, and Security....74
Integrating Assessment Data with Other Systems....75
Guidelines for Data Management....76
Guidelines for Data Storage....76
Guidelines for Data Maintenance, Integrity, and Security....78
Guidelines for Integrating Assessment Data with Learning Systems....82
7. PSYCHOMETRIC AND TECHNICAL QUALITY....84
Background....84
Score Precision, Comparability, and Equating....84
Measuring Change and Growth....85
Validation of Technology-Based Assessments....86
Guidelines for Psychometric and Technical Quality....87
Guidelines for Score Precision, Comparability, and Equating....87
Guideline for Measuring Change and Growth....89
Guidelines for Validation of Technology-Based Assessment....89
8. TEST SECURITY....92
Background....92
Security Threats and Risks....92
Test Security Strategies....93
Guidelines for Technology-Enabled Test Security....95
9. DATA PRIVACY....100
Background....100
Guidelines for Privacy in Technology-Based Assessments....102
10. FAIRNESS AND ACCESSIBILITY....111

Background 111
Accessibility 111
Equity Issues in Technology-Based Assessment 114
Guidelines for Fairness and Accessibility 116
Guidelines for Accessibility 116
Guidelines for Equity in Technology-Based Assessment 119
11. GLOBAL TESTING CONSIDERATIONS..122
Background 122
Translation and Adaptation 122
Availability of Technology Resources 123
Candidate Preparation, Practice, and Orientation to the Technology 125
Guidelines for Global Considerations in Technology-Based Assessments 126
Guidelines for Translation and Adaptation of Technology-Based Assessments 126
Guidelines for Testing Where Technology Resources Are Low 129
Guidelines for Test Taker Preparation, Practice, and Orientation to the Technology..130
PART IV. EMERGING APPLICATIONS IN TECHNOLOGY-BASED ASSESSMENT..........132
Artificial Intelligence 132
Big Data and Social Media 134
Facial Recognition and Analysis 134
Automated Item Generation 135
Regulatory Considerations 136
Conclusion 136
GLOSSARY 138
REFERENCES..147
Appendix: Locked-down Browser Checklist 163

,

PREFACE

The *Guidelines for Technology-Based Assessment* are the result of a multi-year collaboration by the Association of Test Publishers ("ATP") and the International Test Commission ("ITC") to fill a pressing need – to provide guidance and best practices for the design, delivery, scoring, and use of digital assessments, while ensuring the validity, fairness, accessibility, security, and privacy of such assessments. Various other testing standards and guideline documents are available; however, this document is unique in its comprehensive discussion of issues regarding the use of technology in assessment.

This document is divided into four parts. Part I describes the background of, purpose for, and approach to developing the *Guidelines*, and outlines key related documents and references. In Part II, foundational concepts in measurement are discussed, such as validity, fairness, reliability and the need to manage against threats to measurement that may be introduced in technology-based assessments. Part III contains the guidelines, which are divided into 11 chapters, including an introduction and discussion of important considerations followed by guideline statements, which in many cases also include commentary to explain and illustrate the application of the guidelines. Finally, Part IV provides a discussion of emerging applications of technology in assessment that are rapidly evolving and for which the ATP and ITC anticipate best practices and guidelines will be developed in the future.

The *Guidelines* were developed by an impressive team of over 100 authors, technical reviewers, and advisers representing a range of practice areas and regions around the globe. The assessment industry owes these individuals a debt of gratitude for lending their significant efforts and expertise to this initiative (see Acknowledgments). We are grateful for the opportunity to work with these industry professionals and scholars in producing this document.

John Weiner, Chief Science Officer, PSI Services
Stephen Sireci, Distinguished Professor, University of Massachusetts Amherst
Co-chairs and editors
October 2022

ACKNOWLEDGMENTS

The project Steering Committee included co-chairs and editors John Weiner and Stephen Sireci, and committee members Dragos Iliescu, Maria Elena Oliveri, Alex Tong, Alina von Davier, Linda Waters, and April Zenisky.

The following authors contributed to these *Guidelines* by authoring the first draft of the sections listed below and revising these drafts based on comments received from invited reviewers listed below.

Chapter	**Authors**
1. Test Development	
Planning for Technology-Based Assessments	Kimberly Swygert, Wayne Camara
Technology-Enhanced Items	Cynthia Parshall, Belinda Brunner, Kimberly Swygert, Wayne Camara
Gamification, Game-Based Assessment, and Virtual Performance Assessments	Kristen DiCerbo
Universal Test Design	Martha Thurlow
Integrating Assessment with Instruction	Karen Barton
Item Authoring	Joseph Betts
2. Test Design and Assembly	
Linear Test Design	Ye Tong
Adaptive Test Design	Mariana Curi, Nathan Thompson
3. Test Delivery Environments	
Web-Based Delivery	Elizabeth Stone
Offline, Local, and Mobile Delivery	Mark Molenaar
Locked-Down Browsers	Pat Ward
Interoperability	Mark Molenaar, Brandon Dorman
Test Disruptions	Andrew Wiley
4. Scoring	
Automated Scoring of Selected Responses	Ada Woo, Scott Wood
Automated Scoring of Constructed Responses	Susan Lottridge
Technology-Assisted Human Scoring	Catherine McClellan
Scoring in the Event of Technology Disruption and Incomplete Assessments	Sandip Sinharay
Using Item Response Time in Scoring	Dylan Molenaar, Deborah Schnipke
5. Digitally Based Results Reporting	
Results Reporting	April Zenisky
Quality Control in Score Reporting	Avi Allalouf
Maintaining Confidentiality of Score Reporting	Leandro Marino, John Kleeman

Chapter	Authors
6. Data Management	
Data Storage	Steve Polyak, Michelle Barrett
Data Maintenance, Integrity, and Security	John Kleeman, Mark Molenaar
Integrating Assessment Data with Learning Systems	Michelle Barrett, David King
7. Psychometric and Technical Quality	
Score Precision, Comparability, and Equating	Wayne Camara, Jorge Gonzalez, J. Patrick Meyer
Measuring Change and Growth	Yeow Meng Thum
Validation of Technology-Based Assessments	Stephen Sireci
8. Test Security	
Technology-Enabled Test Security	David Foster, Rachel Schoenig
9. Data Privacy	
Privacy in Technology-Based Assessment	John Kleeman, Gary Behrens
10. Fairness and Accessibility	
Accessibility	Mike Russell, Tom Hoffman
Equity Issues in Technology-Based Assessment	Jennifer Randall, Michael Walker
11. Global Testing Considerations	
Translation and Adaptation	Steve Dept
Availability of Technology Resources	Phil Poletti, Derek Trott, Linda Waters
Candidate Preparation, Practice, and Orientation to Technology	Tom Langenfeld
Part IV. Emerging Applications in Technology-Based Assessment	
Artificial Intelligence, Big Data, Social Media, Facial Analytics, Automated Item Generation	Mike Sparling, Ben Taylor, John Weiner

Legal reviewer: Alan Thiemann

Ad hoc reviewers:
Chapter 1. Carol Eckerly, Andreas Orange, Ellen Forte, Rochelle Michel
Chapter 2. Gage Kingsbury, Deb Schnipke, Brent Bridgeman, Rebecca Lipner, Tony Zara
Chapter 3. Joseph Martineau, Brent Bridgeman, John Kleeman, Mark Molenaar
Chapter 4. Joseph Betts, Jill Burstein, EW Looney, Andrew Wiley, Joseph Rios
Chapter 5. John Cronin, Ye Tong, Avi Allalouf
Chapter 6. Mark Reckase, David King, Kristen Huff, Mark Molenaar
Chapter 7. Frank Goldhammer, Michalis Michaelides, Cynthia Parshall
Chapter 8. Camille Thompson, Ray Nicosia, Jim Wollack
Chapter 9. Donna McPartlan, Marc Weinstein, Rory McCorkle, Gary Behrens
Chapter 10. Elizabeth Barker, Marianne Perie, Mark Hakinnen
Chapter 11. Paul Muir, Mike Jodin, Avi Allalouf
Part IV. Alan Thiemann, John Weiner, Ada Woo

Project advisory groups were assembled to obtain input and coordinated review representing assessment practice areas and global regions, led by the following individuals: *Practice areas:* Chad Buckendahl (Education), Craig Mills (Certification and Licensure), Steven Stark (Workplace, Industrial-Organizational), Jennifer Pitkanen (Clinical), David Foster and John Kleeman (Security), Mark Molenaar and Pat Ward (Technology). *Global regions:* Johan Braeken (Europe), Alex Tong and Jimmy de la Torre (Asia), Avi Allalouf (Middle East), and Mariana Curi (South America).

PART I. INTRODUCTION AND BACKGROUND

Purpose of the *Guidelines*

The purposes of the *Guidelines for Technology-Based Assessment* are to provide information about the key factors and issues to consider when designing, delivering, and scoring tests via digital platforms and to provide guidance to test developers, test administrators, and test users on how best to ensure *fair and valid assessment in a digital environment*. The goal of these *Guidelines* is to promote best practices in test development, administration, and scoring to facilitate fair and valid measurement of the knowledge, skills, abilities, and other characteristics (KSAOs) targeted by contemporary assessments used by professionals around the world. As a guidelines document, the purpose is not to specify mandatory practices but rather to inform users about issues and considerations in applying technology-based assessment (TBA). Thus, it may not be possible or necessary for testing agencies and others to adhere to all suggested guidelines in this document.

TBAs comprise a wide range of digitally enabled formats and methods. In these *Guidelines,* any procedure that uses or leverages technology to describe or draw inferences about human characteristics, performance, or predicted outcomes is considered a TBA.

Rationale and Salient Issues

Technology has become an essential part of assessment throughout the testing lifecycle. Test/item design, development, delivery, scoring, reporting, data storage, evaluation, and maintenance are all heavily technology dependent. This is true in education, workplace testing and selection, clinical settings, and professional certification and licensing. Many technology-based applications have become commonplace, such as technology-enhanced items, Internet-based testing, remote online proctoring, data forensics, and biometric measures to authenticate examinees. Emerging trends, such as game-based and gamified assessment, mining "big data" bases, digital social networks, and applications of artificial intelligence and machine learning to devise alternative assessments and procedures, are now pushing the envelope, aspiring to become leading-edge practices.

Regardless of these technological advances, the fundamental concerns with assessment remain the same. It is critical to ensure that the use of technologies in testing adds value through more accurate, accessible, engaging, fair, and secure assessments; without introducing new irrelevant variance in scores or unintended consequences. In other words, TBAs must remain valid for their intended purposes or improve that validity, yielding reliable and meaningful measurement in a manner that minimizes bias. Further, as these new and enhanced technologies increase the global reach of assessment programs, they should be used to facilitate cross-cultural assessment and adaptation.

Early efforts to address issues in the use of technology in testing were put forth by the Association of Test Publishers (ATP) *Guidelines for Computer-Based Testing* (ATP, 2002) and the International Testing

Commission (ITC) *Guidelines for Computer-Based and Internet Delivered Testing* (ITC, 2005). (See also related documents in the References). Since those guidelines were published, many changes have occurred as new technologies emerged and led to dramatic changes in assessment practices. Moreover, assessment developers and users must prepare for new trends on the horizon that signal even more change ahead. Accordingly, ATP and ITC have joined forces to revise and update the *Guidelines for Technology-Based Assessment.*

Scope of the Guidelines

The Table of Contents reveals that these revised *Guidelines* represent a significant update to the previously developed ITC 2005 and ATP 2002 *Guidelines.* The present *Guidelines* address validity and fairness issues in technology-based testing, including how to improve measurement of knowledge, skills, abilities, and other human characteristics, and to address any threats to valid measurement or barriers that the use of technology could introduce. The *Guidelines* specifically address: (a) the planning and design of TBAs), (b) test delivery, (c) psychometric and technical quality issues, (d) security, (e) privacy and confidentiality, (f) fairness and accessibility, (g) integrating assessment and instruction, and (h) global testing considerations (e.g., test translation/adaptation).

Although the scope of these *Guidelines* is considerable, it is important to note the *Guidelines* intentionally avoid duplicating the in-depth guidance provided in other documents pertaining to related topics such as test security, test adaptation, and fundamental issues such as validity, reliability, and fairness. While the *Guidelines* topics address these topics, the focus is on technology enhancements as they pertain to the testing industry. Foundational documents are referenced, where appropriate, to direct the reader to more comprehensive guidance. Furthermore, the *Guidelines* do **NOT** make prescriptions regarding how or when to use technology for testing.

Audience

These Guidelines have been prepared to assist multiple stakeholders in the assessment process. Though not exclusive, the following chart offers suggestions for using this document for diverse audiences interested in TBA.

Suggested Audience	The *Guidelines* May be Useful for
Test developers/ psychometricians	• Describing commonly accepted industry practices to ensure the content of a TBA and the process used to develop it result in a valid and fair assessment for all test takers.
Educational programs	• Informing and enhancing the body of knowledge of computer-based testing and the testing industry.
Public	• Explaining the testing industry processes for determining a test's purpose, the procedures for developing and administering it, and the meaning of its results.
Researchers	• Guiding ongoing research and development of future uses of computer-based testing to enhance the industry.
Technology organizations	• Giving impetus to developing products and processes for continual improvement of TBA.
Test sponsors	• Providing the basics of TBA testing, including development and delivery of tests.
Test users	• Providing information about how to interpret results of technology-based tests and use the results appropriately.
Test takers	• Explaining the testing process, how assessments are developed, what to expect when being administered technology-based tests, and how to interpret results.
Test administrators	• Describing commonly accepted industry practices to ensure the delivery of a test provides a standardized and equitable experience for test takers.
All stakeholders in the testing process	• Presenting quality assurance and quality control procedures to ensure that scores are reasonable.

Development of the Guidelines

In 2018, the International Test Commission ("ITC") and the Association of Test Publishers ("ATP") noted a need to update their respective aforementioned existing guidelines for using technology in testing. The two organizations decided to work together on revised guidelines that would inform the testing communities and lead to better assessment practices.

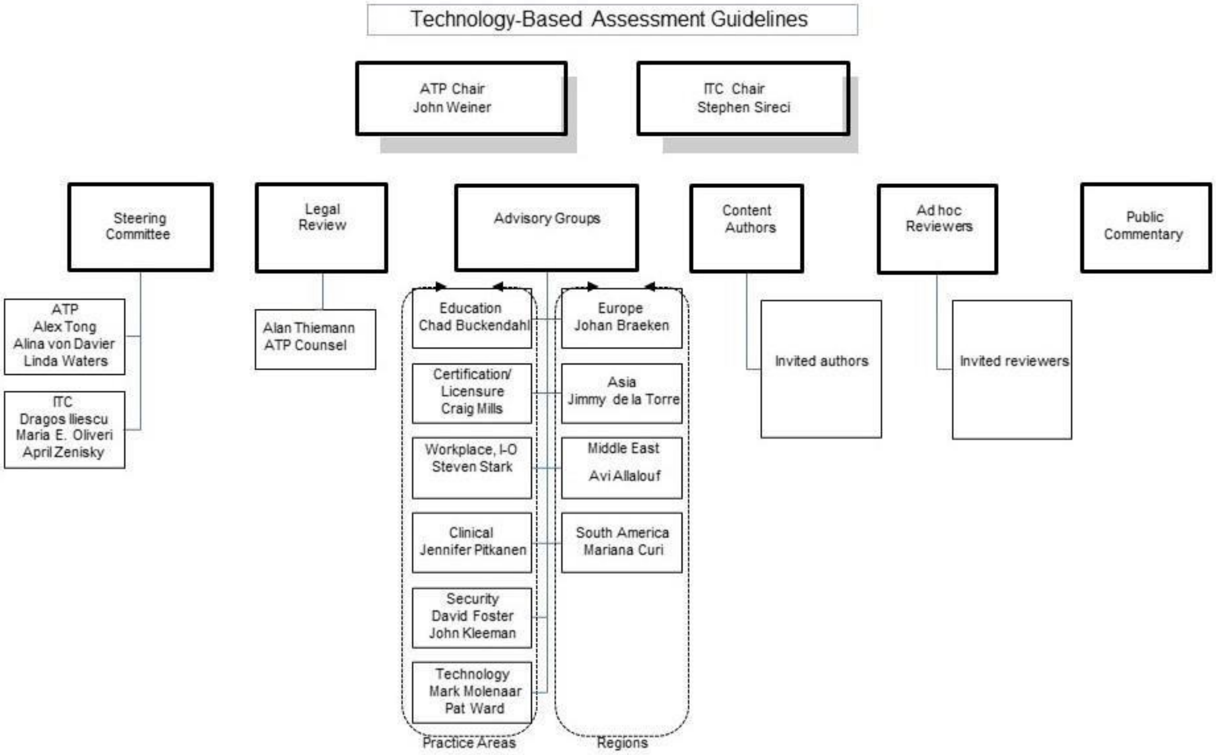
During the ATP annual meeting in San Antonio in February 2018, a joint ATP/ITC meeting was held to secure commitment from the two organizations. At that meeting, John Weiner and Stephen Sireci agreed to co-chair a Steering Committee for the *Guidelines*, with John representing ATP and Stephen representing ITC. Each of the two organizations recruited three members to serve on the Steering Committee. The original three ITC representatives were Kadriye Ercikan, Dragos Iliescu, and April Zenisky. The three ATP representatives were Alina von Davier, Alex Tong, and Linda Waters. Dr. Ercikan took on additional responsibilities that interrupted her participation, so Maria Elena Oliveri stepped in to take her place. Thus, the representation of the Steering Committee was as follows:

<u>Steering Committee</u>

- Co-Chairs: John Weiner (ATP) and Stephen Sireci (ITC)
- ITC Representatives: Dragos Iliescu, April Zenisky, Maria Elena Oliveri
- ATP Representatives: Alina von Davier, Alex Tong, Linda Waters

The Steering Committee's organizational chart shown in Figure 1 indicates how different groups of stakeholders were involved in the *Guidelines* development and revision processes. Brief descriptions of the roles of the Steering Committee, Advisory Groups, and other participants follow.

Figure 1. Guidelines Organization and Process Chart



Co-Chairs: The taskforce co-chairs had overall responsibility in directing all phases of development of the *Guidelines* and served as editors of the entire document.

Steering Committee: The Steering Committee advised in defining the purpose, process, and scope of the *Guidelines* and nominated and approved participants in the Advisory Groups.

Legal Reviewer: The legal reviewer advised on the *Guidelines* to ensure the process followed relevant legal requirements and the referenced salient legal considerations germane to TBA.

Advisory Groups: Advisory group members were solicited to review and provide input on draft documents. A leader for each advisory group practice area and geographic region was appointed to coordinate input from various stakeholders within each group.

Content Authors: A select group of experts was invited to author components of the *Guidelines* in areas of their demonstrated expertise.

Ad Hoc Reviewers: A select group of experts was invited to serve as ad hoc reviewers of the draft *Guidelines* to provide editorial recommendations.

Public Commentary: The development of the *Guidelines* was announced and circulated to industry stakeholder groups for comments/feedback, and the draft Guidelines document has been modified based on public commentary.

Foundational Documents and References

In addition to the preceding relevant guidelines from each organization (i.e., *ATP Guidelines for Computer-Based Testing, 2002; ITC Guidelines for Computer-Based and Internet Delivered Testing, 2005*), the following documents were considered in the development of these guidelines:

- *ATP Privacy in Practice Bulletin Series.* (Association of Test Publishers 2019, 2020, 2021, 2022).
- *Code of Ethics of the American Educational Research Association* (American Educational Research Association, 2011). http://www.aera.net/Portals/38/docs/About_AERA/CodeOfEthics(1).pdf
- *European Union General Data Protection Regulation*. https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en
- *EU General Data Protection Regulation Compliance Guide. (*Association of Test Publishers, 2017).
- *ITC Guidelines for the Large-Scale Assessment of Linguistically and Culturally Diverse Populations*. International Test Commission (2018). https://www.intestcom.org/files/guideline_diverse_populations.pdf
- *ITC Guidelines for Translating and Adapting Tests* (2nd Edition) (International Test Commission (2017). *https://www.intestcom.org/files/guideline_test_adaptation_2ed.pdf*
- *ITC Guidelines on Quality Control in Scoring, Test Analysis, and Reporting of Test Scores*. International Test Commission (2013). https://www.intestcom.org/files/guideline_quality_control.pdf
- *ITC Guidelines on Test Use*. International Test Commission (2013). https://www.intestcom.org/files/guideline_test_use.pdf
- *ITC Guidelines on the Security of Tests, Examinations, and Other Assessments*. International Test Commission (2014).
- *Operational Best Practices for Statewide Large-Scale Assessment Programs* (Council of Chief State School Officers and the Association of Test Publishers, 2013).

- *Privacy Guidance When Using Video in the Testing Industry.* (Association of Test Publishers, 2020).
- *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014).
- *Testing and data integrity in the administration of statewide student assessment programs* (National Council on Measurement in Education, 2012). https://www.ncme.org/publications/new-item

PART II. FOUNDATIONAL CONCEPTS AND CONSIDERATIONS

Validity and Fairness

The *Guidelines* published in this document address many of the fundamental aspects of testing necessary for reliable, valid, and fair assessment. Before presenting those guidelines, this section provides a brief overview of validity and fairness to set the stage for the more comprehensive descriptions of these concepts and their associated guidelines that follow.

The *Standards for Educational and Psychological Testing* developed by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) define validity as "…the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA, APA, & NCME 2014, p. 11). This definition is important because it stresses that validity is not an "inherent" property of a test but rather a judgment that pertains to the use of test scores in a given context.

A concept closely related to validity is fairness. The AERA et al. (2014) *Standards* describe fairness as a "…fundamental issue in protecting test takers and test users in all aspects of testing" (p. 49), and "…responsiveness to individual characteristics and testing contexts so that test scores will yield valid interpretations for intended uses" (p. 50). Essentially, fairness in testing requires test developers to consider the wide diversity of needs and potential inequities within the tested population in all aspects of testing (e.g., test development, developing test preparation materials, test administration, scoring, etc.). In these *Guidelines*, these aspects of fairness are considered in describing uses of technology that promote access to assessments or, conversely, do not inhibit examinees from demonstrating their true proficiencies, attitudes, and other educational and psychological "constructs."

The term "construct" refers to "some postulated attribute of people, assumed to be reflected in test performance" (Cronbach & Meehl, 1955, p. 283). Essentially, the knowledge, skills, abilities, or other attributes measured by a test are called *constructs*. In recent years, technology has helped test developers better measure constructs not amenable to a more traditional testing modality (i.e., paper-and-pencil testing). However, the degree to which technology may change the intended construct to be measured may be a concern in some situations. To understand these issues, we briefly discuss two threats to valid test score interpretations: construct underrepresentation and construct-irrelevant variance.

Construct Underrepresentation/Construct-Irrelevant Variance

While assessments can provide valuable information and insight, they are subject to potential threats to accurate measurement of the constructs they are intended to measure. Messick (1989) summed up these threats as situations where tests either “leave out something that should be included according to the construct theory or else include something that should be left out, or both” (p. 34). The first threat is called *construct underrepresentation*, which means the test is not fully measuring what it intends to measure. Technology can help prevent this problem through innovative formats that address aspects of the construct not possible using traditional item formats such as selected response (e.g., multiple-choice). The second imperfection, which Messick called “construct-irrelevant variance,” (“CIV”) occurs when item or test scores reflect factors the test was not intended to measure. One example is when examinees differ in their proficiency with a computer interface; another is when scrolling on a particular device interrupts a test taker’s reading fluency. Similarly, if examinees take a computerized math test on desktops, their ability to work on a desktop computer may affect their performance to some degree, in addition to their math proficiency. Virtually all chapters provide guidelines related to these issues, particularly Chapters 7 (Psychometric and Technical Quality), 10 (Fairness and Accessibility), and 11 (Global Testing Considerations).

Reliability and Measurement Precision

The scores test takers receive from assessments should be consistent across testing occasions, assuming the test takers have not changed with respect to the proficiencies measured. That is, if test takers repeatedly take an assessment, the scores they receive should provide the same informational value for interpretation and use across repeated administrations of the test. This characteristic of quality in test scores is often referred to as *reliability*, although the AERA et al. (2014) *Standards* expand this terminology to “reliability/precision” to acknowledge how measurement precision is estimated on contemporary tests. The *Standards* define reliability/precision as, “The degree to which test scores for a group of test takers are consistent across repeated applications of a measurement procedure and hence inferred to be dependable and consistent for an individual test taker” (p. 223). Such precision is required for all tests regardless of their use of technology. However, technology-based assessments (TBAs) typically use item response theory in test development and scoring. Thus, the guidelines here address expressing measurement precision using test information functions and conditional standard errors when more traditional estimates of reliability do not apply.

Summary. Our discussion of validity and fairness is brief since there are other important resources on this topic (e.g., AERA et al., 2014; Kane, 2006, 2013; Sireci & Randall, 2021). We also regard consistency of test scoring and measurement precision (i.e., score reliability) as a critical component of quality measurement in TBA, and many of the *Guidelines* speak to issues of measurement precision in TBA. Thus, issues of reliability, validity, and fairness were not only consistently considered throughout the development of these *Guidelines*, but they were also the impetus for us to create them. Our development of these *Guidelines* is intended to inform test developers and users how technology can support more reliable, valid, and fair testing practices; and equally to warn them of issues associated with technology that could interfere with the goals of a testing program. We hope that these *Guidelines* will promote testing practices that leverage and embrace technology and thus lead to more efficient and valid measurement.

Testing Contexts: High-Stakes, Low-Stakes

A key consideration in developing these *Guidelines* is our recognition of the variability of the real-life contexts in which TBA occurs. We aim to ensure the utility of these *Guidelines* across tests and test settings, and in this regard, we particularly acknowledge the differences in stakes associated with different tests. The stakes of a given test result from the consequences placed on the outcome(s) and can vary for different stakeholders even within the same testing context (e.g., instructors, agencies, geographical districts). We note that per the AERA et al. (2014) *Standards*, the higher the stakes for a test (be they technology-based or not), the greater the responsibility of test developers to ensure that evidence supports test quality and the proposed or intended uses (Kane, 2006).

The growing use of technology in testing can increase the extent to which test takers engage with a test due to the implementation of different test item formats and test formats in both high- and low-stakes tests. As noted in the AERA et al. (2014) *Standards,* "Professionals should take into account the purpose of the assessment, the construct being measured, and the capabilities of the test taker when deciding whether technology-based administration of tests should be used" (p. 166). Any use of computer technology for test administration purposes should be thoroughly evaluated for fairness and ease of use within the intended testing population because the extent to which the user interface is readily understood and accessible to test takers will impact their level of engagement (see Chapter 10). When used for test administration, technology can benefit test takers by permitting access to important and practical tools for accessibility and accommodations. However, it can also introduce issues of CIV and call into question the extent to which the data and the results are appropriate for the high- (or low-) stakes use.

PART III. GUIDELINES FOR TECHNOLOGY-BASED ASSESSMENT

1. TEST DEVELOPMENT

Background

The potential for digital technology to enhance assessments has rapidly increased over the last several decades (Bennett, 2015). Innovative items that incorporate multimedia and the assessment of new constructs are two examples of such enhancements. Technology can also help increase test production efficiency, with tools such as automated item generation (AIG) (Gierl & Lai, 2013) deployed to change the standard method of item production. Technology can also be used to integrate assessment with instruction, improve item development, and incorporate universal test design (UTD) principles. In this chapter, we discuss issues related to the development of technology-based assessments (TBAs) and present guidelines in these test development areas.

Planning for Technology-Based Assessments

Anyone planning to incorporate technology into a test, regardless of the stage, should first consider the various manners in which the decision to include technology can directly impact the assessment experience. These ways can be grouped into three broad but interrelated aspects: the *technical*, the *human*, and the *psychometric*.

The *technical aspect* refers to the technology being used to supplement, enhance, or transform an assessment experience. This aspect may be present across the entire assessment experience, including the item development, test design, test delivery, and scoring. Within item development, the use of algorithms, such as those used in the application of cognitive modeling for AIG, can be introduced without necessarily changing the nature of an item or the assessments. However, technology-enhanced items (TEIs) (Sireci & Zenisky, 2016) change the nature of the items. As subsequently described, TEIs can improve the test taker experience, the authenticity of the assessment, or the constructs that can be measured.

The *human aspect* refers to how the introduction of technology may change the test taker experience. All new technology should be introduced to the test taker in a way that avoids introducing construct-irrelevant variance (CIV). Approaches to preventing CIV (Haladyna & Downing, 2004; Messick, 1989) may be as simple as providing an enhanced tutorial to prepare test takers for an item type, such as drag-and-drop, which may require practice. However, these approaches may need to be more substantial, depending on test takers' familiarity with the specific technology involved in an assessment. There may not only be new equipment to use, but also a new approach to testing that requires more interaction on the part of the test taker, or that is more focused on the learning experience, than was common in the past. Human/technology interactions can be complex and unfamiliar. If the technology profoundly changes the nature of the assessment, the test taker will need to become familiar with the equipment as well as the purpose, the scoring model, and other new expectations that go beyond the typical standardized test. In all these cases, the standards of universal design (described later in this chapter) would continue to apply. However, with respect to individuals with disabilities, the technologies used for support and accommodations on the assessment may differ from those they use in the classroom or workplace, and so practice tests and tutorials should be considered to reduce CIV associated with unfamiliar tools and interfaces. New technologies should be introduced to assessments in ways that are congruent with universal design and ensure that additional barriers to access have not been created (see section entitled *Integrated Assessments with Instruction* in this chapter).

Finally, the *psychometric aspect* involves a cohesive analysis and understanding of how technology impacts test scores and the validity evidence to support their use. For example, when a paper-and-pencil fixed form is converted to a computer-based testing format, item performance may change. (This especially tends to occur if a lot of scrolling is involved or additional tools such as online calculators and scratch-paper are provided). However, the nature of the score-based inferences may not be intended to change at all. The psychometric aspect becomes more challenging and complex as the technology more significantly transforms the assessment's nature. The rapid pace of technological developments means the challenges innovative assessment designs presented a decade ago, such as linear-on-the-fly testing, computer-adaptive testing, and multistage testing (Luecht, 2005; Luecht & Sireci, 2011), are being surpassed by the question of what to do with the enormous, diverse sets of data that can now be provided within an assessment. For example, the delivery of a virtual reality (VR)-based surgery simulation may produce so much performance data that simply choosing a logical and defensible scoring method remains a major undertaking (Mislevy et al., 2012).

Technology-Enhanced Items (TEIs)

TEIs can include aspects of technology such as media, interactivity, or response methods that go beyond more traditional assessment methods. TEI complexity can range from the use of technology to supplement item information or response method (e.g., an audio clip within the item stem, hot spot items) to multi-step, integrated tasks or scenarios where technology is used to measure more complex skills.

TEIs may be designed and developed from innovations that pertain to how individual items function and how items interact. Item functionality features may include (but are not limited to) item format (e.g., drag-and-drop, hot spot, simulations), how test takers respond to the item (e.g., voice recognition), the input devices required for responding (e.g., keyboard, mouse, microphone), and stimuli to be delivered with the item (e.g., use of audio, video, etc.). Item interaction features may include (but are not limited to) the use of item sets, how test takers navigate between items, and how particular items are selected for presentation to particular test takers (e.g., within an adaptive exam).

A TEI should be designed to expand an assessment program's ability to measure test takers' knowledge, ability, skills, and other characteristics. Thus, a fundamental goal when including TEIs on an assessment is to use them to counter any existing *construct underrepresentation* by increasing the scope of an exam program's content or construct. For example, TEIs on a language test can be designed to measure listening and speaking skills. TEIs in credentialing exams may also be designed with greater authenticity, whether through realistic software coding tasks, audio clips of heart sounds, or a video clip showing a manager interviewing a potential new hire. In addition to broadening the assessment of existing constructs, TEIs may be included to target higher order thinking skills not previously assessed (ATP & Institute for Credentialing Excellence, 2017).

TEIs may also be included in an assessment to increase student engagement (Huff & Sireci, 2001). Greater engagement can increase student motivation and effort, which can contribute positively to the validity of the test scores (Wise, 2015). Additionally, it can improve the face validity of the exam, which can enhance stakeholders' (e.g., test takers, employers) perceptions of the value of the assessment and the resulting qualifications. However, claims of increased engagement require validation.

TEIs should strive to avoid the presence of CIV in test scores. Simply substituting a TEI for a more familiar item type can contribute CIV if test takers are not given time and opportunity to become familiar with the novel technology before testing. A test taker's low score on the item may indicate a lack of familiarity with the technology, rather than a lack of the knowledge, skill, or ability construct being assessed. There are additional risks when technology is added to an item. For example, if audio is added to a test item, the performance of an item becomes dependent on correctly functioning headphones. If video is added to test items, the file size of the video clips may require additional bandwidth for test delivery. Audio and video features may also differentially affect test takers with disabilities. For example, if audio is added to a test item, performance becomes dependent on the availability of a text transcript for test takers who are deaf or hard of hearing. If video is added to test items, performance becomes dependent on closed captions for test takers who are deaf or hard of hearing and is dependent on audio descriptors for test takers who are blind or have low functional vision.

These training and technical issues should be investigated to ensure CIV does not affect test scores. Similarly, each new TEI should be examined to verify it does not introduce bias for any subgroup of test takers (e.g., gender, ethnicity, test takers with disabilities, country of education; see Chapter 10). In other instances, some TEIs may require more exam time for a test taker to respond than a traditional item type. In this case, the time limits or item counts may need to be adjusted.

TEIs can be easier to remember than traditional items and are sometimes worth multiple points, and thus may have a potentially greater impact on test taker scores. If a test administration schedule is long, there is a risk that test takers who test late in the window may gain prior knowledge about the content of a TEI and benefit from this prior knowledge.

This brief overview indicates technology can enhance items and tests in important ways. However, TEIs must be designed, developed, and evaluated carefully to fulfill that vision of improved measurement. The use of "technology for the sake of technology" should be avoided despite the appeal of emerging technology that may seem especially engaging in assessment. Research will always be needed to ensure a successful TEI, with studies likely to include cognitive labs, usability studies, accessibility analyses, and comparability research. Technical issues will need to be considered, and these are likely to include delivery platforms, bandwidth, and security. Collaboration with experts in technology system user experience (UX) and user interface (UI) design is important to the successful design of TEIs.

Gamification, Game-Based Assessment, and Virtual Performance Assessments

Game-based assessment (GBA), gamification, and virtual performance assessments (VPAs) are relatively nascent endeavors compared to the maturity of the field of assessment overall. As such, their development may involve experimentation and revision in design phases that may stretch the comfort level of many assessment developers. Thus, the costs in terms of budget and time of creating GBA and VPAs (Andrews-Todd et al., 2021) can be high, or at least higher than traditional assessment types. Also, the use of process or activity stream data and techniques for analysis are evolving. Technology systems must be built to accommodate the storage and efficient querying and retrieval of these data (and this storage may require test taker consent under some privacy laws). Formerly distinct fields, including educational data mining and psychometrics, must combine their practices to develop ways to turn these data into inferences and insights. Given the emergent nature of the practice, standards themselves are emergent, so the guidelines presented here are based on practical experience.

The engagement and motivation that games produce are attractive to those who struggle to ensure test takers' performances represent best, or even good, effort. Games can be engaging and promote complete absorption in an activity by using well-calibrated challenges and motivating objectives. In addition, games can provide rich, novel environments in which students can apply their skills. Also, the application of newly acquired knowledge to new contexts is often the ultimate goal of learning. Games allow test takers to interact in ways that traditional assessments do not, allowing more than binary correctness of response as evidence of learning. The advent of digital environments means it is possible to gather problem-solving information from test takers as they engage in activities and use it to make inferences about what they know and can do without requiring them to stop and take a traditional test. Incorporation of game designers in task and item design can lead to better tasks and items as long as the designers are fully informed about validity goals. Others have sought to borrow particular elements of games and use them in assessments to increase motivation without designing a full game experience, a practice termed “gamification.”

GBAs and VPAs should be considered for constructs that are otherwise difficult to measure and contexts in which motivation is a concern. Games provide a wide variety of types of environments and interactions, which enables gathering multiple types of evidence. For example, when assessing systems thinking, games can immerse players in a system and provide choice points woven into game play that evaluates their understanding of that system. These assessments should also be considered for situations where there is a desire to combine learning and assessment in the same experience. Note, however, GBAs, VPAs, and aspects of gamification may not be designed to be accessible to assistive technology users and so developers should follow the Web Content Accessibility Guidelines (https://www.w3.org/WAI/).

Universal Test Design

Universal Design is an approach to design and development that originated in the field of architecture (Story, Mueller, & Mace, 1998) to ensure access to and use of buildings and physical devices for all individuals, including those with sensory and physical disabilities. Since its first appearance, its application to other fields has grown. One of many areas in which universal design principles have been applied is assessment design, development, administration, and interpretation (e.g., Ketterlin-Geller, Johnstone, & Thurlow, 2015; Thompson, Thurlow, & Malouf, 2004). This application to educational assessment has a complementary parallel in the emergence of Universal Design for Learning (UDL, e.g., Rose, Meyer, & Hitchcock, 2005).

The AERA et al. (2014) *Standards* define UTD as "An approach to assessment development that attempts to maximize the accessibility of a test for all of its intended test takers" (p. 225). Other organizations provided similar definitions. For example, the *Operational Best Practices* (CCSSO & ATP, 2013) defined UTD as "a set of construction principles that seeks to maximize the accessibility of an assessment for all students by developing items and content without distractions or irrelevancies" (p. 216). As described by the National Center on Educational Outcomes (NCEO; 2022), "The goal of applying universal design principles to assessments is to be able to design and develop assessments that allow participation of the widest range of students, and result in valid inferences about their performance."

UTD is a means for creating accessible assessments for all test takers, including special populations (e.g., individuals with disabilities, multilingual learners) that benefit **all** individuals, not just those with disabilities or language learning needs or individuals from different age, gender, or cultural groups. It applies to technology-based assessments (TBAs) and paper-and-pencil tests, and its principles are recognized and applied in many countries (Hayes, Turnbull, & Moran, 2018). Elements of a universally designed assessment include the removal of barriers that are irrelevant to the construct being assessed. Thus, it is critically important to first define the target skills for the construct being assessed and the construct-relevant and irrelevant skills needed to access the assessment (Ketterlin-Geller, 2008). See Chapter 10 for further discussion of technology accessibility considerations and guidelines.

A central tenet of universal design is to consider diverse users from the onset of test development in an attempt to reduce reliance on retrofits, which provide solutions with limited effectiveness and can compromise original design. To address this, a validity framework leveraging UDL principles has been developed as a means for writing guidelines to both reduce CIV inherent in task design, as well as reduce the introduction of new sources of CIV (Dolan et al., 2013). In addition, accessibility needs with these types of assessments will be best addressed by having UI/UX experts working closely with content and measurement experts (see the Web Accessibility Initiative: https://www.w3.org/WAI/).

Integrating Assessments with Instruction

The best-known approach to assessment integration with learning is likely formative assessment (Black & Wiliam, 1998), which involves gathering evidence about student learning and adjusting instruction accordingly. Such integration should reflect a system of evidence collection and communication that directly relates to the goals and outcomes for learning that are provided to and used by both teachers and students (Council of Chief State School Officers [CCSSO], Formative Assessment for Students and Teachers State Collaborative on Assessment and Student Standards [FAST SCASS; CCSSO, 2019]).

As with any valid assessment design, assessments integrated with learning should reflect their purpose and use--to inform and drive instruction and learning, where both the teacher and student are engaged and empowered to act based on the results, and where learning is not assumed to be static. Ensuring the validity of TBAs integrated with learning thus requires an intentional and continual focus on how the assessment design considers the context of administrations in learning environments, valid interpretations, and appropriate decisions and use relative to the intended purpose. Integrating learning content into the assessment using adaptive technology can be a valuable component of assessments designed to be integrated with instruction.

Context is uniquely vital to the design of integrated assessments, given the variability of teachers, students, and learning environments. The volume of information to learn, and the pace and modality of both instruction and learning, can vary across time, teachers, students, and subject areas. With seemingly infinite variations in learning contexts, the design of valid TBAs that integrate with such dynamic situations can present complex challenges not typical of other standalone assessments. For example, when assessments take place in classrooms as part of a learning program, there is often greater flexibility and less standardization in administration conditions and can be particularly challenging when the assessments are technology based. There can be very real logistical burdens for the administration of TBAs on students and teachers, such as limited access to technology, technology training (and training resources), and constraints on class schedules limiting dedicated administration time. The design challenge for integrated assessments is to balance standardization intended to maximize trust in the results (reduce measurement error, increase reliability and validity) with the reality of the unstandardized manner in which they may be administered.

Given the complexity and diversity of learning environments and administration contexts, assessments integrated with learning should be part of a system of complementary assessments that are ultimately useful and valid for dynamic learning. The National Research Council (2001) made these recommendations, which are still relevant today: Assessment systems should be *comprehensive,* with a range of approaches and types of measures so students can truly demonstrate what they know; *coherent,* such that models of learning are connected across both instruction and assessments; and *continuous,* to capture and demonstrate progress over time. A well-designed system of integrated TBAs is an appropriate solution for developing flexibly and validly designed assessments that are comprehensive in coverage, coherent with and reflecting instructional models for a given program, and supportive of ongoing opportunities for collecting and providing feedback about performance evidence (OECD, 2020).

Therefore, a valid system of TBAs integrated with learning should be designed with intentionality of purpose and use. The design of this system should also consider and attend to varied contexts of dynamic learning and provide helpful evidence-based information for students and teachers. TBAs integrated with learning are not limited to K-12 schools; they can also be included in credentialing contexts such as longitudinal continuing certification programs in medicine and other professional practice areas. Other guidelines related to integrating assessments with instruction can be found in Chapter 6 (Data Management).

Item Authoring

Item authoring is a fundamental aspect of test development. Strong item-authoring processes can lead to high-quality items, which are a necessary part of a validity argument that supports the interpretation of the test scores. Item authoring revolves around item writing and review. Traditionally such activities have involved empaneling a group of subject matter experts (SMEs) to carry out the necessary tasks. However, remote work is an option for accomplishing these same tasks. This change in the traditional panel's structure has some important considerations unique to conducting work in a distributed and remote manner. The guidelines on item authoring intend to help guide the transition from more traditional in-person item author panels to the use of digital technology and remote tools for completing this work.

A first consideration when using technology to facilitate item authoring (e.g., item banking applications that allow for secure group review of the same items) is managing a remote group's workload, assignments, and interactions. It may be important to check in regularly on member productivity when working remotely. Managing an in-person group gives individuals the visual context for verbalizations of others, whereas managing a remote group may require mitigating misunderstandings that evolve due to the lack of physical context. Video conferencing applications that allow members to interact can help with this issue. It may be helpful to have both a method of interacting as a group and private communications channels outside of the group to ensure quality group interactions. Of course, confidentiality laws will apply to any video or other recordings of SMEs (see Chapters 6 and 9).

Allowing remote proctoring to monitor video and audio feeds while listeners focus on screen and keyboard activity can help proctors observe and control the environment and alert test developers to potential risks and threats. For instance, if a keyboard listener recognizes a keystroke, e.g., 'print screen,' or a combination of keystrokes, e.g., Windows key + Shift + S, it can alert the proctor and/or facilitators of the potential for content theft. As AI becomes more embedded in proctoring applications, there is the future potential to mitigate risks in remote authoring contexts.

Guidelines for Test Development

Guidelines for Planning a Technology-Based Assessment

1.1 TBA development plans should include definitions and descriptions of the use of technology and its impact on measurement properties and non-psychometric features, including:

(a) expected impact of technology on the *test purpose*.

Comments: If a testing program transitions from a paper-based format to a TBA, the reasons for the change, such as improved validity via construct representation, efficiency (e.g., shorter testing time or immediate score reporting), and fairness (reduction of CIV), should be explicitly described to all stakeholders. In addition, test results should be evaluated to assure technology did not introduce unexpected effects.

(b) intended changes that will affect the psychometric properties of the test (e.g., measurement precision, comparability of scores).

(c) intended changes to test characteristics that are not psychometric in nature (e.g., reducing cost or increasing accessibility to the test).

(d) how technology will impact the assessment of the *constructs of interest.*

Comments: If technology is used to improve an existing testing program, plans should clearly indicate whether the intent is to measure the same constructs measured by the older testing program, whether new constructs will be measured, or whether the technology will be used to measure the same construct in an improved way. An explanation of the expected gains from the changes should be provided, and a validity research agenda to evaluate whether the expected gains are realized should be planned. If the assessment will measure new constructs, an explanation of the reasons for measuring these new constructs should be provided (e.g., an emerging competency in the field of practice demands new assessments). Any aspects that may be expected to impact test security (positively or negatively) should be made explicit, as these can impact validity arguments.

(e) costs of the technology weighed against the expected benefits.

Comments: These costs may include the initial costs of technology investment, costs to retrain item writers and test developers, costs to increase or change the size of the item pool, costs to inform and retrain test takers to use the new technology, costs of licensing to use the technology, and costs to educate other stakeholders such as test score users.

(f) how the additional technology should inform *item development*.

Comments: Information technology experts should fully evaluate the software used to develop the items. Resources should be devoted to retraining/recruiting item writers for the new method.

(g) how the additional technology should inform *test design and delivery*.

Comments: The potential impact of test taker access to new technology should be fully evaluated (e.g., consider the impact of enhancements that require Internet access for test takers in areas where broadband access may be inconsistent-or not available -see chapter 11). Also, alternate design solutions should be considered. And where relevant, descriptions should also be provided regarding how technology can support access to the test for test takers with disabilities and multilingual learners.

(h) how the additional technology is expected to impact *scoring*.

Comments: The impact of new technology for scoring (e.g., automated scoring of essay items) and new data for scoring (e.g., multiple responses to a problem-solving item) should be evaluated. Steps should be taken to reduce the potential for scoring to seem like a "black box" (e.g., test takers can take many actions but do not know how their actions are being scored). Any potential biases in scoring technology should be properly evaluated and tested (see Chapter 10).

1.2 Planning for TBAs should include studies to survey test takers about their experience with the new technology, including usability, efficiency, and any technological failures or mishaps.

Comments: Data should be collected to ensure usability or access issues did not impact test taker performance (e.g., analysis of omit rates or item display failures for a new item type). Efforts should be made to reduce potential confusion on the part of the test taker in using the new technology, to minimize the introduction CIV.

1.3 Planning for TBAs should identify groups of test takers who may be differentially impacted by technology to identify and minimize the introduction of CIV.

Comments: Experience with technology is likely to interact with culture, disability, socioeconomic status, and other test taker characteristics. Testing organizations should study the diversity of the test taker population to identify and address any of these interactions and plan accordingly (Sireci, 2020). Similarly, UTD should be included to prevent the introduction of obstacles for test takers with disabilities or access challenges.

1.4 Planning for TBAs should include the design of tutorials for navigation of test elements.

Comments: It is important to provide tutorials for test takers to become familiar with the testing system user interface to avoid unintended effects on test scores (See Chapter 11, Candidate Preparation, Practice and Orientation to the Technology).

Guidelines for Technology-Enhanced Items

1.5 The development of TEIs should begin with an analysis of construct needs.

Comments: Construct needs analysis should be based on the testing organization's goals and requirements as reflected in the existing test specifications, content blueprint, and/or skills map. The analysis should involve SMEs representing important characteristics of the target testing population and score-user population--and consider test taker diversity (Randall, 2021). The use of TEIs may be reserved for those constructs or cognitive processes for which it is difficult to assess the depth and/or breadth of the construct with more traditional item types. The result of the construct needs analysis should be a listing of the specific areas of construct or cognition that the technically enhanced items are targeted to address.

1.6 The specific innovation/type of TEI selected for a program should support psychometrically sound assessment while minimizing construct-irrelevant factors.

Comments: TEIs should be designed not to require knowledge, skills, or abilities irrelevant to the construct being assessed. TEIs should maintain test fairness and avoid any new group differences.

1.7 The exploration of TEI innovations should include a cost-benefit analysis that estimates the costs associated with the innovations weighed against the potential benefits derived from implementation.

Comments: The cost-benefit analysis should begin early in the process of implementing TEIs and should consider not only financial costs, but other less tangible costs related to the development and implementation of the TEIs. These considerations include ongoing content development efforts, scalability of the innovations, the capability of the item to provide multiple measurement opportunities, the efficiency of the item (i.e., the time it might typically take a test taker to respond versus the assessment information provided through the item), the complexity of the UI, technical difficulty in delivering the item (e.g., large file sizes in low bandwidth situations), and speededness.

1.8 An evaluation of the feasibility of TEIs should be conducted using prototypes, where these may involve SMEs, test developers, technical experts, and software developers.

Comments: Information gathered from feasibility evaluations should be used to revise prototypes and select those most promising for further consideration. TEIs can be iteratively refined through SME judgments and user-centered research (e.g., think-aloud protocols, cognitive labs) to ensure the intended functionality. Usability studies may help ensure the *tested population will understand and be able to use the item interface, so no CIV interferes with measurement. The result of iterative refinement should be a prototype that has been adequately specified and includes consideration of instructions needed to respond to the item, the physical layout of the item on the screen, how to score the item, requirements related to assets (e.g., images, video) to be delivered with the item, and any directions necessary as to the format of the item prompt, scenario, and response options. (See also 1.12.)*

1.9 Item analyses such as differential item functioning analyses should be applied to identify potential CIV in TEIs.

Comments: The degree to which TEIs may interact with test taker characteristics such as gender and race/ethnicity should be studied both quantitatively and qualitatively to ensure TEIs are appropriate for all test takers. Psychometric modeling in general, such as item and person fit analyses, should be fully leveraged to evaluate CIV associated with TEIs.

1.10 Procedures should be developed to address workflow management and storage of TEIs and associated ancillary materials.

Comments: Workflow management systems for TEIs should address media requirements, complex response strings, procedures to address workflow management and storage outside of an existing item-authoring system (e.g., how media will be stored securely outside of the system, how a TEI will be stored within an existing item bank, additional needs for test takers with disabilities, etc.).

1.11 Item-writing guidelines should address the technical requirements of the TEI to encourage production of consistent, high-quality items.

Comments: TEIs may be supported with item-writing guidelines that instruct the item writer to specify the graphical file types supported in the exam, define the response marker displayed with the item, and denote likely incorrect regions, along with the key region, on the item image. Audio items may have item-writing guidelines that include specifications for the minimum and maximum length of any audio clip to be included, a limit on the number of speakers within a single audio clip, and specifications for the types of item content that may be provided in audio form.

1.12 Pilot testing of TEIs should include testing technical requirements with the appropriate delivery platforms to confirm the intended rendering and verify the TEI is functioning as intended.

Comments: The purpose of the pilot test is to gather item statistics that will provide information on how the TEI performs and as a final indication for operational use. Pilot testing should specify the data collection method, sample size, and characteristics desired to generalize findings to the intended population, the analyses to be conducted, and the criteria required to determine whether elements need further iteration and testing. If multiple devices or platforms are used, studies should ensure the TEIs operate similarly and CIV is not introduced. For example, a study of interface *differences, such as screen size, or the amount of scrolling needed, can reveal effects on item timing and/or performance.*

1.13 When feasible, the pilot test for the TEI should be administered under the same testing conditions that will apply when the TEI is delivered operationally.

Comments: An alternative approach may be needed if TEIs cannot be piloted within the existing test form. In these instances, a special pilot administration may be required, with special effort to obtain motivated responses.

1.14 Tutorials, practice items, and other communications should be developed so test takers will have the opportunity to familiarize themselves with the TEIs before testing.

Comments: Communication to all stakeholders (e.g., test takers, educators, employers, parents, and the public) about TEIs should be available in advance of the launch of TEIs to provide the opportunity to prepare for them. Tutorials should include information about the test (e.g., number of items, timing, types of items), the testing procedures (e.g., how to navigate through the system, how to exit), and how to respond (including how to change a response). Information to be communicated before the initial administration of a new TEI should include the test blueprint areas addressed by the new TEI, how the TEIs look and function, and how the TEIs will be scored.

1.15 Clear and sufficient on-screen instructions regarding how to interact with a TEI should be provided during the test on each item screen.

1.16 Some TEIs may need additions or adaptations to address accessibility needs.

Comments: For example, to meet the needs of deaf and hard-of-hearing candidates, a video clip that includes speech may need to be delivered with captioning or a text transcript. See section 10.1 for guidance on test accommodations.

1.17 TEIs should be designed to reduce the impact of memorability.

Guidelines for Development of Game-Based and Virtual Performance Assessment

1.18 The appropriateness of GBA or gamification of an assessment should be evaluated for suitability for the assessment's purposes.

Comments: Evidence supporting the validity, reliability, and fairness of GBAs is nascent and developing. GBA may be sufficient to support low-stakes decisions but may not be sufficient to support decisions in high-stakes environments.

1.19 The degree to which different groups of test takers may interact differently with features of the GBA or gamified features of the assessment should be studied.

Comments: When studying test taker interactions with assessments, it is recommended to notify *test takers of the intended use of the data for such research purposes and that their data will be anonymized and aggregated (see Chapter 9).*

1.20 When gamification is used in TBA, collaborative teams of test and game designers should work together to ensure testing purposes are being met.

Comments: Collaborative game-based test development requires establishing common ways of working together across teams and collaborative design sessions. Games, assessment, and content experts use different terminology and processes. Thus, time is needed to establish common ground. People who have created traditional assessments may need to think differently about authoring and evidence.

1.21 A principled design process should be used in GBA and VPA, including models of skills, task design, and evidence derived from the tasks.

Comments: While important for all assessment, the complexity of GBA requires building the chain of evidentiary reasoning. Use research-based skill models where possible to identify or create a student (learner) model. Learning progressions make excellent models, as a progression's stages can become game levels.

1.22 When designing games or simulations, game play should be targeted to the constructs of interest and should cover all intended aspects of the construct required to make the intended construct inferences. The impact of experience with game play should be mitigated in the design.

Comments: Merge game design and assessment design practices. Consider which game genres are consistent with the type of activity the target construct suggests. Also, consider which game mechanics align with the types of activity needed to generate assessment evidence.

1.23 When used to assess learning, game loops should be linked to the skills to be learned.

Comments: The game loop actions players engage in repeatedly should be tied to the knowledge and skills to be learned or assessed.

1.24 When designing gamification, a range of game mechanics should be considered.

Comments: When implementing game features in an assessment, the design should consider more than just scoring "points." Games are attractive for many other reasons, such as quests, narrative, collaboration, and challenges. Non-desirable and inauthentic potential test taker behaviors (e.g., taking extreme risks just to see what happens) should also be considered.

1.25 The potential negative effects of competition on test scores should be avoided, particularly those related to reinforcing negative stereotypes.

1.26 Game design tools should be used to build early wireframes, storyboards, and level and game descriptions.

1.27 "Play testing" should be used early in the GBA development process to revise and improve the assessment and gather evidence of the skills and knowledge players use to advance in the game.

Comments: Play testing can be helpful in test design. Such testing could involve recruiting a small number of individuals in the target demographic, observing them interacting with the prototype, having them think out loud while engaging in the test, and noting the knowledge and skills they use to complete the activities.

1.28 Alpha and beta testing should be used to gather data to evaluate GBAs and VPAs.

Comments: Ideally, alpha and beta tests should involve sufficient numbers of test takers to provide stable estimates of item and test statistics. However, user feedback will be especially valuable. The user interface and experience should confirm that the interface does not inhibit learners. Collect enough information from alpha and beta testing to evaluate your evidence aggregation models (e.g., item response theory (IRT), diagnostic classification models, BayesNets). Allow sufficient time in the schedule between alpha and beta tests to make revisions.

1.29 Design of assessment tasks and scoring rubrics should focus on features of the performance rather than right and wrong or dichotomous scores, including presence or absence of actions, counts of actions, and sequence of behaviors, all within the context of the game (i.e., which actions were taken at what place in the game).

1.30 Exploratory data analysis and data mining can be used to verify hypotheses about the construct of interest and identify other game evidence that may improve inferences.

Comments: The evidence available in GBA flows from the game mechanics built into the design. It may be helpful to develop initial hypotheses about what actions in the game are validity evidence during the authoring stages. Multidimensional IRT, diagnostic classification models, and Bayesian networks are all possible forms of evidence aggregation.

1.31 Consider alternate indices of measurement precision where traditional estimates of test score reliability may be less appropriate.

Comments: Test-retest reliability estimates should be interpreted with caution if it is likely that players will learn about the construct as they are playing the game. Generalizability Theory studies may be more appropriate if players all play the same scenarios and produce the same evidence or if it is possible to manipulate certain components in an A/B trial setup to build evidence. It may be helpful to review internal structure and beware of level effects in which data gathered from the same level may share common covariance (similar to testlet effects). These effects can be statistically modeled.

1.32 Reporting of assessment results should make explicit how actions in the game or performance tasks relate to constructs of interest.

Guidelines for Universal Test Design in Technology-Based Assessment

1.33 UTD principles should be embedded in the description of the constructs measured at the design and development phases and during administration.

1.34 The access needs of the population of individuals to be tested should be identified in the design.

1.35 UTD principles should be applied to test administration to enhance access for the broadest range of the target population without disrupting the construct-relevant aspects of each item and the assessment as a whole.

1.36 Test administrators should be trained in universal design principles to ensure access for all students during test administration.

1.37 Access needs should be continually evaluated to ensure the TBAs address them.

Comments: Several studies could be conducted to evaluate the effectiveness of specific UTD applications. These studies include cognitive lab studies with individuals in the target population and statistical procedures, adjusted for small populations if needed, to evaluate the differential performance of subpopulations within the population to be assessed.

Guidelines for Developing Technology-Based Assessments Integrated with Instruction

1.38 Prior to design, a theory of action should be developed for the system of TBAs to be integrated with learning.

Comments: In developing an assessment/instruction theory of action, specify the decisions to be made from the results and who should make them. The anticipated evidence relevant and necessary for those decisions should be well documented. In addition, clearly describe the TBA design elements, components, and processes, including considerations for administration and providing feedback in learning contexts, as well as a defined process to ensure quality, accuracy, and validity across stakeholders and end users. The context of assessment and instruction should be considered to ensure assessment information is consistent with the intended uses as specified in the theory of action. The theory of action should specify the end users of the information, including learners, to ensure the assessment design provides actionable information. The theory of action should specify empowering learners to be involved in and responsible for their own learning.

1.39 Where appropriate, software tools and data management system(s) should support the development of assessment content that can be tagged with metadata important for learning.

Comments: TBAs that go beyond providing evidence of what a test taker knows and provide information regarding why a test taker does not do well on an item or task (e.g., evidence of misuse of strategies, misconceptions, misunderstandings, errors, etc.) will be more helpful for instructional purposes. The system should provide information regarding actions teachers can take in response to the evidence. (e.g., What will the instructor consider next if a learner does well on the item/task? If a learner does not do well on the item/task, what specific misunderstanding might the learner have?). Program level metadata useful for instruction, learning, and validity evaluation should be considered, verified/validated, and documented to support decision-making and use. Metadata can also be used to provide immediate and relevant feedback directly in the system to drive student learning, including redirecting, scaffolding, or correcting misunderstandings/misconceptions.

1.40 Where appropriate, item presentation and response modes should be diverse to provide meaningful, non-redundant information and reflect instruction pace, depth, and design.

Comments: TBAs can reflect various item and task types used in instruction, which may better reflect how instruction and learning occur. The assessment should offer a range of response methods that best align with the targeted concepts and relative complexity.

1.41 When adaptive TBAs are used, reported information should include what the test measures for specific individuals or groups of test takers.

Comments: Information should be provided to learners and instructors about the concepts measured, how the items/tasks contribute to a domain of learning and an overall score (where relevant), and the degree to which the items represent the intended knowledge, skills, and abilities. If the assessment is intended to drive "personalized learning," the results should integrate into an instructional environment that supports personalized instruction where the range and pace of learning and instruction varies. If results will be used in a more traditional learning environment (classroom, not self-paced), where instruction is targeted and delivered within a grade-specific classroom, it may be helpful to adapt the assessments deeper into the grade rather than merely across grades (Barton, 2020).

1.42 Consider building instructional examples within reporting to reflect evidence and support instructional decision-making when possible.

Comments: It may be helpful to provide examples of what the items measure and examples of learner work that represent both mastery and misunderstandings.

1.43 Integration testing (i.e., how the assessments connect to and work alongside learning environments) should be conducted early in assessment development, rather than solely at the time just before or after deployment.

Comments: Specific TBA features such as the ability for instructors and learners to set goals, determine criteria for success, and track performance should be an essential part of integration testing.

1.44 If a recommender system is leveraged, ensure assessment designers or content experts evaluate the validity of the process the engine uses.

Comments: Evaluate the data considered and algorithm constraints invoked to select recommended content. Conduct studies to ensure assessment engines are efficient, accurately connected to data, and that the resulting recommendations are useful. See also Chapter 7 (Validation of TBAs).

1.45 If a recommender system is leveraged, ensure assessment designers or content experts evaluate the validity of the process the engine uses.

Comments: Evaluate the data considered and algorithm constraints invoked to select recommended content. Conduct studies to ensure assessment engines are efficient, accurately connected to data, and that the resulting recommendations are useful. See also Chapter 7 (Validation of TBAs).

Guidelines for Item Authoring and Review

1.46 When item authoring is conducted remotely, ensure remote participants have enough Internet bandwidth and other technology resource requirements to allow for efficient working capacity in a secure environment.

Comments: Ensuring Internet bandwidth is vital as members will potentially be using audio and video over the Internet along with any web-based applications for completing their work. Degraded connectivity can cause issues during the panel and inhibit productivity. Use trusted encryption or other security methods (e.g., secure file transfers, secure virtual private networks, etc.) to keep transferred information secure. Provide a specifications sheet outlining technology requirements to ensure panelists have the necessary resources loaded on their home-based work machines. Ensure all panelists have the correct digital screen configurations and hardware specifications. Technology experts should be available to troubleshoot if issues arise. Have a backup plan to continue the facilitation of the panel should disruptions occur to the connectivity or video conferencing platform being used. A secure environment also requires adhering to appropriate privacy laws.

1.47 Authoring systems should provide a useful workflow for item-authoring panel members and content managers to create and manage item content in various stages.

Comments: An effective item-authoring system will likely include monitors and alerts that inform facilitators when items have been written, ensure team members understand where items are in the workflow, and alert participants when new work is available for access and interaction. Also, ensure panel member status in the system is changed after each event from the access and permissions needed for the event to non-access status.

1.48 When using remote item-authoring panels, support group-related work with a group interaction method that enables members to interact with each other (e.g., group video conferencing).

Comments: It may be helpful to have a private channel available for facilitators to interact with item authors. In such cases, video conferencing and chat functions should be private and outside the communications with others in the group. This arrangement might be needed to provide specific feedback to an individual that should not be privy to others in the group. Allow members to give feedback on what works well and what is problematic to help improve the process in the future. Use audio/video and screen-sharing applications to allow for a face-to-face interactive engagement that emulates the type of discussion and focus on work product that one would see in an in-person panel and use chat or text features to provide written feedback when needed.

1.49 When using an Internet-enabled content management system, configure as many item writers' guidelines as possible by default in the system.

Comments: For example, if some aspects of an item must have at least a certain number of characters or a specific number of options/keys, etc., configure the system to enforce as many of those business rules as possible. Ensure item writers have access to item-writing guidelines, style guides, and needed reference materials while completing their work.

1.50 Privileges for participants in the development and management of test content should be set to meet the participants' specific needs (e.g., item writer, review panel, program administrator).

Comments: For instance, sensitivity review panel members might need read-only access if they are allowed to do their first reviews on an individual basis and make ratings or remarks. However, an editorial review by content editors would need to have read/write/update privileges. In addition, confidentiality/nondisclosure agreements should be established with all personnel who have access to items and test forms or other sensitive information (test takers' PI and test scores).

2. TEST DESIGN AND ASSEMBLY

Background

The use of technology enables advanced test design and assembly models that are made scalable through computer automation of test construction methods, utilizing item banking and delivery systems and software. This chapter outlines guidelines and best practices for assembling linear and adaptive tests, highlighting considerations to ensure valid and fair assessment. Overviews of the most common test designs are provided, followed by guidelines for implementing these designs effectively.

Linear Test Design

Linear test design refers to a fixed test form administered to test takers. The linear test is assembled either a priori, or on the fly, but it is fixed, and the items will not change or update once the candidate starts taking the test.

There are two dominant types of linear tests. One type is the linear test that is built ahead of time, published, and is available for a certain number of candidates for a period of time. This type of test will be referred to as fixed-form testing (FFT). The other type of linear test is called linear-on-the-fly testing (LOFT) (Folk & Smith, 2002; Stocking, Smith, & Swanson, 2000). LOFT forms are also fixed in length, but they are built “on the fly,” drawing from a pool of pre-calibrated items such that each test taker receives a combination of test items that is not exactly the same as other examinees. LOFT forms are fixed once the test taker begins the test.

For both FFT and LOFT, classical test theory (CCT) or item response theory (IRT) can be used to calibrate test items and assemble equivalent forms (Gibson & Weiner, 1998). While FFT can readily be delivered either on paper or via computer, LOFT is delivered via computer using algorithms to build the test “on the fly.” Rather than constructing a large number of FFTs, a carefully curated item pool is made for LOFT assembly. The same test specifications are followed for each LOFT form generated to ensure fairness for all candidates.

FFT and LOFT are both effective test designs under the most common standardized testing conditions. Test developers should choose a design that fits best with the purpose and use of the assessment. FFT can be an effective design for both measurement and cost considerations when test administrations are event-based and less frequent, or the volume of test takers tends to be small. LOFT can be advantageous when the testing window is long or on-demand, when the volume of test takers is high, and when there are security concerns, as in high-stakes testing.

Security is another consideration in the use of FFT and LOFT. The test delivery system may randomize item ordering and answer choice order to enhance FFT security. Additionally, some testing organizations may opt to have many FFT tests available during the same administration window. During that time, individuals taking a test will be administered one of the many available forms. On the other hand, LOFT uses dynamic forms generation software algorithms to assemble a unique combination of test items to comprise an equivalent test for any given test taker. This provides more security as receiving a unique exam form makes memorizing and sharing exam content difficult. Another advantage of LOFT is that pilot questions can be rotated to gather sufficient data while minimizing exposure before operational use.

Adaptive Test Design

Adaptive testing, often called computer-adaptive testing, refers to the personalized delivery of assessments to test takers with optimized precision in estimating ability. The personalization can occur in at least one of two ways. The test can adapt the number of items, whereby some test takers experience shorter/longer tests than others. It can also adapt the nature of the items, typically by matching item difficulty to examinee ability. Adaptation can also be based on machine learning models, cognitive diagnostic models, or other approaches. Adaptive tests offer several significant advantages over traditional linear testing, leading to much shorter tests, while increasing score precision, fairness, test security, and test taker engagement.

An adaptive test consists of several components (Kingsbury & Weiss, 1984; Luecht, 2016):

- large and fully calibrated item bank where all items have complete and accurate metadata;
- starting point or initial ability estimate;
- item selection algorithm;
- scoring algorithm; and
- termination criterion (or criteria).

Additional sub-algorithms are often added, such as introducing content or exposure constraints to the item selection algorithm, but the general approach remains the same. These components and parameters serve as the basis for design of an adaptive test and should be documented.

In general, there are two types of adaptive test designs: item-level and multistage. Item-level designs adapt to the test taker after every item. Multistage-adaptive designs adapt after pre-designated blocks of two or more items (sometimes referred to as "testlets" or "modules"). There are advantages and disadvantages of item-level and multistage-adaptive designs. For example, item-level designs may have increased measurement precision. In contrast, multistage designs may allow test takers to review items within a stage and provide more balanced use of the item pool (Luecht & Sireci, 2011).

The stakes of the test are one of the primary drivers of adaptive test designs. As with all tests, sufficient measurement precision and content coverage are essential criteria to be considered when weighing the potential advantages and disadvantages of adaptive test designs. Potential advantages include reductions in testing time and costs (for organizations that pay for seat time), improved test taker engagement, and increased precision and test security. Potential disadvantages include inability for test takers to review previous answers, nonuniform item pool usage, and vulnerability of the adaptive algorithm to manipulation by test takers.

Test Assembly

Considerations for test assembly are similar to those for the test designs discussed in this chapter. The tests will be assembled to meet a series of constraints such as content outlines, timing considerations, desired statistical characteristics (Classical, IRT, or other), and item exposure controls. Performance-based assessments may be similarly assembled in accordance with specifications for a simulated activity, scoring rubric, and other elements. For higher-stakes and some tests designed for formative purposes, test forms will be assembled from a bank of items with known statistical properties derived from pre-testing. The test assembly constraints are designed to produce equivalent score interpretations across multiple forms of the test. Automated test assembly software is often used to produce multiple forms of linear tests that meet the same constraints. Adaptive item selection methods also assemble tests based on these constraints but incorporate an additional factor in test assembly--the test taker's performance on previously answered questions or groups of questions. Information explaining the use of AI systems or automated software for content generation purposes is found in Part IV.

Guidelines for Test Design and Assembly

2.1 Design and development of a technology-based assessment (TBA) should take into consideration important factors related to the purpose, content, and psychometric characteristics of the assessment as used in a digital environment.

Comments: Technical and practical considerations include but are not limited to content domain representation, item types, testing volume (annual number of tests administered), the psychometric model used to calibrate items and score tests (e.g., CTT versus IRT), the size of the available item pool, and the costs associated with developing a sufficient pool of items or performance-based assessment.

2.2 TBAs should be built to the content and statistical specifications of the test blueprints. If multiple linear forms are being administered, they should be parallel.

Comments: Parallel forms are equivalent in psychometric properties, including content coverage, cognitive complexity, and difficulty. Content equivalence can be ensured via item selection software, including algorithms that require fulfillment of test specifications (see definition of "algorithm" in the Glossary, which provides the distinction between software automation and AI). *Random forms may be sufficient for certain low-stakes scenarios (e.g., practice). Total test time and other practical considerations may be relevant here as well.*

2.3 A field-testing design should be developed, and items should be field-tested and calibrated with an appropriate model (e.g., IRT or another model) with sample size thresholds sufficient for stable item parameter estimates.

Comments: Issues to consider in the field-test design include security of the items and ensuring test takers are composed and motivated as representative of the intended population. Also, consider avoiding cueing operational items and ensuring adequate time for test takers to complete the items. A common approach for LOFT is to embed experimental (unscored) items in an operational section of a test form. Post equating is more common for FF exams, wherein a final equivalent form is selected after the test administration.

2.4 **Statistical analysis should be carried out at both the test and item levels to support test form development.**

Comments: Such analyses may include item parameter drift studies to ensure the items remain stable across administrations, timing analysis to assess speededness, item psychometric properties, differential item functioning analyses, and test form psychometric properties.

2.5 The item bank should be evaluated routinely to inform test assembly, maintain security, and plan for future item development.

Comments: Such evaluations may include item exposure and usage, capacity to yield test events that are aligned with assessment targets as defined by the test blueprint, depth, availability of items in the existing item bank, security threats, testing volume, and other factors.

2.6 **When FFT is used for an assessment program, multiple linear fixed forms should be developed when possible and as needed to manage content exposure.**

Comments: It is important to consider testing volume, content exposure, retest policies, and security threats in planning the number of alternate forms.

2.7 When using LOFT for an assessment program, consider the item bank composition, its ability to support the LOFT model, and rules that will govern or constrain test assembly. These would include the amount of form overlap, masking of field-test items, content domain representation, and handling of accommodations.

Comments: Before the test administration, simulations ensure the item pool will be able to support robust, parallel, and reliable tests for each potential test taker. It is important for assessment organizations/programs to evaluate and consider establishing rules regarding the overlap between item pools for adjacent administration windows for security reasons. During the administration, if certain items need to be masked (no longer used operationally), simulation may be necessary to ensure LOFT continues to work with the revised item pool and that no additional bias in item selection is introduced. The size and representation of the item bank across content domains and statistical properties are essential considerations to support a reliable and efficient LOFT design.

2.8 Development and implementation of an adaptive test include many choices regarding test design parameters, and these choices should be researched and documented for validity and defensibility purposes.

Comments: Such choices include item bank size, IRT model, distributions of item parameters, item selection method, exposure constraints, content constraints, scoring method, and termination criteria. Characteristics of the item pool (e.g., size, parameter distribution) should be investigated early in the process to determine what is necessary to meet the desired measurement properties of the test. These test properties may include score standard error, average test length (for variable-length CAT), item bank utilization, content domain coverage, and item exposure rates. These choices, and their reasons/research, contribute significantly to validity, for example, documentation that shows that content coverage is achieved with CAT.

2.9 CAT design should be informed by simulation studies to investigate how a final version of the adaptive test would perform under various situations. Simulation studies should be designed to support the goals of the CAT program (e.g., producing much shorter exams or producing more precise scores).

Comments: All independent variables in such a study should be realistic, to the extent possible. Dependent variables of the simulation studies should reflect the results of interest. If the test is multistage, the simulation should reflect the statistical properties of the testlets (modules) and the testlet selection criteria that will be used. Before implementation, a CAT item bank should be investigated for needed exposure constraints and appropriate termination criteria to inform final decisions. It is also important to investigate correct module (multistage) or item assignment as well as analysis of how often items/modules are used.

2.10 Time limits should be based on an empirically derived threshold rather than an arbitrarily selected one. Additional research and consideration should be given for examinees who require extra time.

Comments: Time limits can be informed by analyzing item response times from field-test or operational data.

2.11 A published TBA should include appropriate documentation for technical stakeholders (testing experts, regulators, lawyers) and non-technical stakeholders (e.g., test takers, parents, supervisors).

Comments: For technical audiences, the design, development, and validation evidence for TBAs should be documented in accordance with industry standards and best practices; see Chapter 7. Psychometric and Technical Quality. For automated test construction methods such as LOFT and CAT, results of the simulation studies and the test development process are typically documented in the technical report, including the parameters used and why they were selected. For non-technical *stakeholders, it is helpful to provide an overview and explain what to expect when taking the exam, how the adaptive model works (if used), and how scores are reported.*

2.12 The software platform used to deliver the test should be fully capable of meeting the technical and practical needs of adaptive testing, including CAT and multistage test algorithms, item selection software use of IRT data, and technology requirements for fast and reliable implementation, while minimizing introducing construct-irrelevant variance.

Comments: IRT item calibrations may be calculated within the system or externally and imported into the testing software platform for use in CAT. Potential latency should be considered in evaluating the system capability.

3. TEST DELIVERY ENVIRONMENTS

Background

There are many environments in which technology-based assessments (TBAs) can be administered. These environments include web-based, offline, local, mobile, and locked-down delivery systems. All these environments call for careful consideration of interoperability issues and potential test-taking disruptions. This section discusses issues and guidelines associated with TBA delivery environments.

Web-Based Delivery

Web-based test delivery – also referred to as "Internet testing, Internet-based testing, or online testing" (Foster, 2016, p. 36) – affords flexibility in administering assessments and capturing and relaying data between repositories and other points in the testing system. Although testing centers also use Internet-based services for receiving and delivering tests, the opportunity to expand assessment administration beyond dedicated testing centers can enhance the scalability of computer-based testing. However, web-based delivery also brings with it the need for additional consideration of efficient, appropriate, and secure data structures, repositories, and transmission methods (Luecht, 2016). Increasing test complexity also increases the need to evaluate system capacity. Adaptive tests (Chapter 2) and technologically enhanced item and task types (Chapter 1) put additional demands on delivery and data transfer, especially when that transfer occurs in real time over the Internet. In adaptive tests, the system requires input from the test taker (e.g., an item response) to be actively captured and used to make an on-time decision (e.g., selecting the next item). With technologically enhanced content, bandwidth issues may arise from transferring large files (e.g., audio or video components or high-resolution graphics) or supporting interactive elements.

As one example, web-based delivery is increasingly employed in K–12 learning and testing in the United States. The finding that many schools and districts do not have adequate technological infrastructure and bandwidth has led to efforts to improve the quality and availability of Internet access in schools (e.g., Fox & Jones, 2019). Another example is in remote clinical assessment that increased as a result of the Covid-19 pandemic (Wright et al., 2020; Society for Personality Assessment, 2022). Similarly, there are efforts to expand Internet availability across the globe. Therefore, inequities in access to robust and sound technology must be considered in developing and delivering web-based assessments. A robust and smoothly functioning web-based delivery environment can provide a secure assessment session while mitigating lags and other test-taking delays that may introduce construct-irrelevant variance and demotivate test takers.

In addition to guidelines for web-based assessment delivery, this chapter includes guidelines focusing on interoperability and test-taking disruptions. Additional relevant chapters in this document address security (Chapter 8), data privacy and confidentiality regulations (Chapter 9), and accessibility for individuals with disabilities and other special needs (Chapter 10). As is true in the other sections, web-based delivery guidelines should be considered in the context of the stakes of the assessment, the level of supervision or proctoring of the test taker, and the specific laws and regulations of jurisdictions that apply to particular tests or types of testing. Protecting secure test material and test taker rights must be prioritized, regardless.

Offline, Local, and Mobile Delivery

Availability and continuity of the test-taking experience are crucial in TBA, particularly in high-stakes testing. Ideally, testing should be able to continue under the most challenging circumstances, such as power outages, Internet unavailability, local network congestion, and device issues. Depending on the nature and stakes of the assessment, different solutions (e.g., alternate testing modalities) can be implemented to mitigate risks related to the availability and continuity of the assessment, varying from entirely web-based to computerized to mobile delivery. Offline and local delivery are considered specialized options of computerized delivery, where "the test content is downloaded in its entirety before the beginning of the test administration event" (Foster, p. 236).

Offline test delivery occurs when test content is installed or downloaded to the individual device, after which the test can be administered without any network connectivity. In local delivery, local infrastructure, such as a local server in a school or test center (or use of a USB device where the test is downloaded and cached in the computer), is leveraged to store and serve test content, eliminating the dependency on live Internet connections during the test-taking process but still requiring a (local) network connection. Hybrid models are also potentially possible, combining aspects of web-based, computerized, and mobile delivery (e.g., a local computerized infrastructure, which receives occasional content updates over the Internet a couple of days before a major testing event).

These different delivery options have benefits and challenges concerning functionality, bandwidth, connectivity, implementation, security, and other logistics. Regarding functionality, restrictions in the use of external sources (e.g., YouTube videos), item types (e.g., online simulations), and test designs (e.g., item-by-item adaptive testing leveraging software or an application programming interface [API] and item bank) are influenced by the testing purpose and practical factors (e.g., item and other test development resources, stakeholder perceptions, construct representation, testing time). As for bandwidth/connectivity, some solutions only require connectivity up front to download complete test packages. Other web-based options cache some items in advance, but still require some connectivity during the test for the computer/tablet. Solutions with local storage typically require a synchronization mechanism and exchanging test content, results, and administrative data, which can be harder to implement. Solutions should also address security and privacy, and confidentiality. Storing (encrypted) assessment content and results on a local network or device poses security risks such as potential data tampering and manipulated synchronization, as data could be remotely stored for an extended time. In addition, the effort to deploy and manage a solution with offline capabilities can pose a significant burden on (local) administrators required to install software, synchronize data, prepare workstations, and other tasks.

Mobile delivery provides additional opportunities and challenges in end-user experience and validity, requiring accounting for various form factors, screen sizes, and input types (Wools, et al., 2019). Finally, bring-your-own-device policies and remote proctoring options, allowing testing from home computers or other non-managed devices, pose additional challenges in the areas of lockdown (detailed below).

The guidelines in this chapter provide recommendations for implementing a robust test-taking experience. This section should be read in conjunction with Chapter 6 on Data Management, Chapter 8 on Test Security, Chapter 9 on Data Privacy, and Chapter 10 on Fairness and Accessibility.

Locked-Down Browsers

In the late 1990s and early 2000s, computer-based testing began migrating from proprietary applications for displaying questions to systems that use an underlying HTML rendering engine. Like many of the changes in the Internet revolution, the advantages were obvious: more powerful display capabilities, standard formatting controls, and cross-platform support. Today, almost all test delivery software uses HTML rendering engines to display questions, including test center or desktop-based systems that are not connected to servers or the Internet.

Online testing opened new opportunities and quickly grew to be a significant method for test delivery. One big challenge with online testing was test security: the browsers used for everyday browsing to popular websites are not acceptable mechanisms to deliver a test. Testing companies created the locked-down browser to address this problem. A locked-down browser delivers online content on a full screen, securing the environment to prevent a test taker from accessing other sites and engaging in test fraud. These secure browsers are used across the testing spectrum: in testing centers, classrooms, and at home. Most remote proctoring systems have an integrated locked-down browser, and some companies use them for personnel training, item banking, and item reviews.

Locked-down browsers are just one key piece in the greater security discussion. They have three primary functions: full-screen display, preventing access to non-authorized web-accessed sites and digital tools, and preventing content from being stolen. The locked-down browser displays the content on a full screen, typically with browser features such as a hidden address bar. Users may or may not be restricted from switching to other applications on the computer or device. In addition to seeing a full-screen display of the testing application, users will be prevented from accessing non-authorized tools such as email, the Internet, or messaging. A locked-down browser may allow the test taker to access tools such as a calculator or allow limited browsing of external sites. Of course, any browser does not prevent the user from accessing external resources, such as books, paper documents, or other devices. Locked-down browsers aim to prevent content exposure, visibility, or theft.

It is essential to recognize that a locked-down browser does not prevent all forms of test fraud. Methods for bypassing locked-down browser security range from a simple hidden camera in the room to a more sophisticated attack (e.g., where the secure browser is running in an undetectable virtual machine [VM] window on a host computer). Locked-down browsers create barriers but do not prevent all cheating and content theft methods.

Interoperability

Interoperability is the ability of technology systems to communicate with one another through an agreed-upon set of minimum shared information fields and in an agreed-upon format. While there are many specification-setting bodies, these *Guidelines* cover only fundamental interoperability requirements for creating and exchanging accurate data. Being able to accurately allow test takers to test on a wide variety of devices and platforms is necessary with the internationally recognized interoperability specifications available today. Systems that make (assessment) data interoperable can more easily exchange data, prevent vendor lock-in, protect investments (in content creation and data collection) and allow for a multi-vendor, best-of-breed ecosystem--as outlined by Educause (Brown, Dehoney, & Millichap, 2015) on the Next Generation Digital Learning Environment.

There are many open specifications and standards, some targeting specific countries, regions, cultures, or (sub) industries, and new standards are likely to arise during the lifetime of this document. In addition to providing guidelines on technology standards for exchanging (assessment) content, metadata, and statistics; the interoperability guidelines in this section provide an overview of good practice, specifications, accessibility requirements, and digital credentials and pathways by standards organizations such as Advanced Distributed Learning Initiative (ADL, 2019), Access 4 Learning (A4L) Community, Aviation Industry CBT Committee (AICC), Dublin Core Metadata Initiative (DCMI), HR Open Standards, IMS Global, the IEEE Learning Technology Standards Committee (LTSC), Common Education Data Standards (CEDS) and the Ed-Fi Alliance. The interoperability guidelines in this section provide an overview of the most relevant and universally applicable guidelines today and are not intended to comprise an exhaustive list.

Test Disruptions

For many testing programs, one of the most damaging events is a significant disruption during test administration. Unfortunately, significant testing disruptions have been common, with examples coming from both the education (e.g., the state of Florida in 2015) and the credentialing (e.g., the Canadian CPAs in 2019) communities. Disruptions have included test takers being unable to enter the system to begin their test,[1] being thrown out of the system while completing the test,[2] and being unable to access critical reference material required to complete the test.[3] These disruptions may come about due to simple misfortune or poor planning by the testing bodies or their service providers. Still, disruptions can also result from deliberate attacks from individuals or groups with malicious intentions. The guidelines on this topic aim toward reducing and eliminating test disruptions. These disruptions are in addition to other test irregularities that are not technical in nature (e.g., behavioral disruptions, outside

[1] https://abovethelaw.com/2018/03/bar-exam-software-debacle-causes-testing-delays-across-the-country/

[2] https://www.ajc.com/news/local-education/statewide-internet-outage-disrupts-delays-georgia-milestones-tests/U48hiQ9SviZ1rvFBw1cJxK/

[3] https://business.financialpost.com/news/fp-street/further-delays-add-up-to-major-frustration-for-8000-would-be-accountants-after-testing-snafu.

distractions, such as sirens or construction noise), which, while important to ensuring the integrity of the test administration process, are out of scope for this Chapter.

There are two key components that are important for every testing organization to consider for minimizing test disruptions. First, testing organizations should systematically identify the risks of testing disruptions associated with their specific program. For each risk, organizations should create a list of activities designed to mitigate this risk, including identification of likely internal or external disruptions in test events, so that the organization and its vendors have an agreed plan for reporting/resolving these disruptions. These activities should include the *proactive* collection of data and information that can be used to monitor the test administration and identify areas that may be experiencing issues in real time as the events occur. Second, testing organizations should develop with service providers (e.g., test delivery, proctoring) a comprehensive *response plan* that can be followed if a testing disruption occurs.

It should be noted the guidelines in this section are intended to be generic and not as specific as what will be needed for an individual testing program. While some programs may need to determine procedures that will fit with a remote proctoring model, other programs may only have testing in schools or testing centers. Readers are advised to review the other chapters in this section for additional information on different administration models.

Guidelines for Test Delivery Environments

Guidelines for Web-Based Delivery

3.1 The test delivery system should support the secure exchange of test material and test taker data as appropriate for the testing purpose.

Comments: The inclusion of technologically intensive test elements should be limited to those required to support accessibility and to support making valid inferences about the constructs measured. Identify dead zones and other areas of lower bandwidth capacity that may not be suitable for testing. If the expected architecture is not sufficient, consider options such as boosting the wireless signal, throttling the bandwidth to prioritize usage by test takers, planning administrations at non-peak usage times, staggering administration to limit the number of simultaneous test takers, or switching from events/windows to on-demand testing to reduce the number of concurrent test takers.

3.2 Technological requirements of the assessment system should be provided in advance (e.g., bandwidth required per individual test taker, hardware, and software needs), and scalability of system resources should be commensurate with the number of test takers.

Comments: This includes the appropriate balancing of test taker load on the system across system resources and capitalizing on the geographic proximity of test takers to servers. Consider buffer size, capture rates, and temporary data storage (e.g., in the cloud or required locally) for securely *storing and retrieving data in case of an Internet outage. Include system redundancies (failover solutions) to prevent disruptions to test taker sessions. Using a specialized content delivery network is recommended when test content leverages extensive rich media such as streaming video.*

3.3 Web-based delivery systems should be designed to prevent the loss of test taker response data (e.g., capturing each test taker response when submitted, where possible).

3.4 Extraneous or unwanted computer functions should be disabled (e.g., sticky keys or other accessibility settings that may be inadvertently triggered and impede the test administration).

Comments: One option would be to use technology such as a locked-down browser to prevent individuals from engaging in activities disruptive to the test session (e.g., locking out key combinations) or that may negatively impact test security. Functions the test taker requires or prefers may be exceptions to this guideline.

3.5 Access to the test system should be provided in levels commensurate with the minimum required for specific roles (e.g., system administrator, proctor, test taker, in decreasing order of privilege) to allow adequate control of the testing scenario and limit test security risks.

Comments: Use systems architecture aligned with the security level required to ensure the integrity, availability, privacy, and authenticity of data transmitted and received.

3.6 Scoring should be conducted at the server level to prevent compromise or subversion at the browser level (or device/hard disk level), where possible.

3.7 The test delivery system should be evaluated and confirmed to be accurate before operational testing, including opportunities for users to engage with the system.

*Comments: It is vital that firewalls and other security measures (e.g., pop-up blockers) not impede key aspects of the test from being administered. The system as a whole must be tested before operational use to identify interoperability or other issues***.** *Provide tools for administrators to run system and connection tests in advance to ensure everything is set up correctly before the test administration and to troubleshoot any problems during test administration.*

3.8 System performance should be monitored throughout testing.

3.9 The test system/platform should be capable of allowing test takers to resume the test (where they stopped or as close to it as possible) after a service disruption or a planned break.

Comments: The test delivery system may not capture all test taker responses until a task is finished and so it is likely test takers who experience disruptions may need to restart at the last point at which their responses were recorded by the system. The spirit of this guideline is to credit test takers with as much of the test they completed as possible to minimize the extent of the disruption on their time and mental energy.

3.10 Appropriate security protocols should be implemented to restrict access to the Internet/wireless network and prevent hacking and data theft.

3.11 Troubleshooting information should be provided to test takers and support staff in a timely and appropriate fashion for addressing technical issues and errors that arise during administration.

Comments: This information should include contact information (e.g., hotline) for real-time technical support and logging and escalation of delivery issues. Also, the testing organization should provide a guide to the delivery system that details the typical causes of errors and their manifestations and the actions a test taker can perform or that can be handled remotely, which should promote better communication between test developers, administrators, remote proctors, and test takers, as well as provide more timely and efficient intervention.

Operationally, the system should capture error messages and detailed information necessary to diagnose administration errors and facilitate data recovery across multiple sessions (e.g., for an individual whose test session terminates prematurely and must restart or resume the test). Error messages should be designed to be informative and factual without being unnecessarily alarming or raising security concerns. If message security is an issue, one option is to provide an error number the test taker can relay to help desk staff, who can troubleshoot accordingly.

3.12 Contact staff should be trained to answer routine questions and escalate those requiring more technical assistance. Points of contact should be able to facilitate the resolution of test delivery issues efficiently and effectively.

3.13 Ethical issues surrounding the impact of negative feedback should be taken into consideration, and directions for accessing support should be provided where possible.

Comments: Aspects for which this consideration is important include but are not limited to the language used to convey incorrectness of test taker responses, score interpretation guidelines, and psychological assessments that evaluate personality or job fit.

3.14 Testing should be conducted under appropriate environmental conditions, and outcomes should be interpreted in light of those conditions.

Comments: Provide guidelines on required testing conditions (testing workspace, lack of distractions, Internet capability, computer specifications, etc.). These conditions should be optimal for test takers and consider practical issues, such as taking breaks. Provide appropriate guidance for the level of supervision required, which will be dependent on the testing stakes and context. In the case of unproctored, self-administered testing in a low-stakes environment, guidance should be provided to the test taker detailing required test-taking conditions and procedures.

3.15 Guidance should be provided regarding the required level of authentication of test taker identity.

Comments: This information should be provided clearly and in advance of the testing event, so test takers understand what is expected. The appropriate level of authentication will depend on the nature/stakes of the exam. Security (Chapter 8) and Data Privacy (Chapter 9) should also be considered with respect to the collection and storage of test takers' personal information (PI).

3.16 Testing procedures should be monitored to ensure that security is maintained (e.g., through in-person proctoring or supervision, or by using remote monitoring through cameras if the stakes of the assessment warrant).

Comments: Test takers should be provided with accurate information in advance regarding the type of monitoring to be used, including information legally required in the relevant jurisdiction. This information may be provided within a test taker agreement or exam procedures document, given to and agreed by the test taker in advance of test administration.

Guidelines for Offline, Local, and Mobile Delivery

3.17 Test delivery systems (whether offline, local, or mobile) should be robust and secure, including capabilities for graceful degradation, encryption, auditing, and meaningful system messaging.

Comments: Graceful degradation of systems permits test sessions to continue as long as critical functionality required to take the test is not impacted (i.e., to fulfill all requirements to result in accurate scoring/reporting).

(a) Encryption of data should be used in transit and at rest when dealing with confidential test content, test taker data, test results, and administrative data, as well as for all personally identifiable data where privacy requirements exist (see Chapter 9).

(b) An auditing system should be in place to keep track of all actions performed by all actors (test takers, proctors, administrators) and test disruptions logged by systems to replay events and non-technical irregularities when required. Recording of testing events for individual test takers must comply with applicable national laws and regulations.

Comments: The actual conditions in which the test was delivered should be (anonymously) recorded for analysis and validation purposes.

(c) Meaningful messaging should be provided to all end users in case of incidents such as severe system failure.

Comments: See also 3.10.

3.18 Web-based delivery methods should allow for (central/cloud) availability and temporary drops in (local) Internet connectivity.

(a) Central (cloud-based) systems should (automatically) provide for failover and scale up system resources when the number of concurrent end users increases.

(b) **System resources should be scaled up when the number of concurrent end users increases.**

(c) In case of system failure, new instances should be spun up automatically and take over required tasks without impacting ongoing end-user sessions.

(d) End users should be evenly distributed across available resources to handle increased loads.

(e) Test takers (traffic) should be distributed across multiple (cloud provider) availability zones and regions to ensure responsiveness and short loading times when testing across multiple geographic locations.

Comments: Specialized Distributed Denial of Service prevention is recommended for high-profile testing programs and events.

(f) Disconnected testing sessions should be able to continue during temporary drops in (local) connectivity or network congestion.

Comments: Test content and results should be configurable to be cached to allow for testing to continue during temporary drops. Test content could be cached up to X number of items up front, configurable based on exact testing program requirements and stakes unless the testing format prohibits this (e.g., in computer-adaptive testing [CAT]). Computerized delivery methods should download test content up front and allow for testing without a dependency on (live) Internet connectivity during the administration event. Offline options should leverage the individual device to allow test content delivery during the administration event.

(g) If a temporary drop in connection turns out to be permanent, an end user (e.g., administrator, proctor) should have the option to securely close the session.

Comments: Where possible, test results should be preserved locally in a secure and end-user-friendly fashion. Checks need to be put in place to prevent manipulated synchronization of data out of sight for an extended time (e.g., repeated uploads of results). Ideally, the program can take measures to securely close the session (delete the cached items in the browser). Administrators and other service providers must not be allowed to have unauthorized access to, or capability to change, the responses of any test taker.

3.19 Where software installations are necessary, the solution should be easy for an end user to install (e.g., downloading the test content should be straightforward).

Comments: The solution should check whether sufficient disk space is available for test content and results. In some cases, the offline option should be able to run or be booted from a USB stick or other removable media. It should be easy to connect to the local server by local clients, such as workstations, to access test content and persist results.

3.20 Local test delivery systems should allow for storing test content, results, and administrative data on local infrastructure, such as a local server.

Comments: The local server should be easy for a local administrator to install and configure. It also should be able to leverage the local network for serving test content, administrative data, and receiving results. Security issues (Chapters 6 and 8) would need to be addressed.

3.21 Mobile delivery methods should allow for test taking on the go, not depending on permanent Internet connectivity.

Comments: Test content may be (partially) downloadable to the device, when appropriate (e.g., when encryption and other security measures are taken according to the stakes of the test), allowing for uninterrupted test taking on the go while the end user is moving and (temporarily) out of an Internet connection.

(a) The testing interface and test content should be rendered responsively based on the form factor, orientation (portrait, landscape) of the device, and available screen estate, unless the testing program and stakes prohibit this.

Comments: In some cases, the allowed classes of devices could be limited to offer a comparable testing condition to all end users. This could be implemented by either white- or black-listing of devices classes, form factors, accessibility supports, input types (keyboard, mouse, touchscreen), orientation (portrait, landscape), operating systems, and browser versions.

(b) Available device capabilities should be leveraged where applicable, such as (external) keyboards, mouse input, touchscreen, and stylus usage.

3.22 All stakeholders should thoroughly prepare for the testing administration event.

Comments: Some stakeholders could hold multiple roles (e.g., an assessing organization could also be a platform vendor and test center provider). Vendors should provide reasonable (fallback) options applicable to the administration event: online, offline, local, and/or mobile delivery, and should perform thorough quality assurance on all delivery methods and combinations thereof on a wide range of devices and conditions, including technical benchmarks and stress tests on central (cloud) infrastructure in representative conditions before the testing event.

(a) **To prepare for test administration, test delivery vendors should ensure the exposure of actual test content will be as limited as possible (including minimizing test content available on servers and (automatically) expiring/removing content). They should also provide diagnostic tools, documentation, and technical training to assessing organizations and local staff.**

(b) **End-user tools, documentation, and training for local staff to support the testing event properly should be provided.**

(c) **Access to representative practice materials should be provided to schools/test centers and end users, including test takers, to familiarize them with the system and testing content on applicable domains, and with the system's functionality.**

(d) **Testing location staff should run diagnostics on workstations/devices and local servers (if applicable) before the testing event.**

(e) **Test administrators and other service providers with a need to access the system (e.g., proctoring services), should receive tools, documentation, and instructions to prepare for the test.**

Guidelines for Locked-Down Browsers

3.23 Locked-down browsers should prevent access to non-authorized tools.

(a) **Locked-down browsers should display content in a full window, hiding all other applications, taskbars, and other operating system features, including clocks, applications, network access, and sound controls.**

(b) **Locked-down browsers should detect when running in a hosted VM window.**

Comments: Virtual machine (VM) software is the most common attack vector for a locked-down browser. When a locked-down browser is running in a hosted VM window, the window is locked down, but users can access email, messaging, and browsing in the host computer, allowing cheating and content theft. Detecting some VM software applications is extremely difficult. These VMs, marketed for privacy and piracy, focus on spoofing/fooling an application so that it does not know it is running in a VM.

(c) **Locked-down browsers should detect and block any remote desktop access.**

Comments: Remote desktop access is a technique that allows a person on another computer to view and/or control the screen of the test taker’s computer. Most locked-down browser attacks involve a VM used to host remote access.

(d) Applications running on the test delivery system that are not associated with the test being delivered should be blocked from showing while the test is in progress.

Comments: All unauthorized applications must be blocked from running during a testing event.

(e) Locked-down browsers should support allow/deny lists of the sites that can be visited or blocked, ensuring that only approved content is shown.

Comments: Locked-down browsers should support browsing to approved external domains while at the same time blocking access to unapproved external domains

(f) Locked-down browsers should provide a method for test delivery applications to ensure that the locked-down browser is running.

Comments: One attack vector to bypass test security is running the test outside the secure browser. Locked-down browsers must provide a secure mechanism for test delivery software to validate that the secure browser is running. The most common method is embedding and verifying an identifier in the agent string; however, an attacker can easily spoof this. A more secure approach is for the locked-down browser to provide a function that takes a random string parameter and returns an encrypted version of that string to the server for verification and validation.

(g) Locked-down browsers should block examinees from utilizing a multi-monitor configuration to bypass security.

Comments: Users often have multiple monitors running simultaneously. The locked-down browser should be capable of detecting these configurations and ensuring that no applications or content is shown on the monitors that are not displaying test content. Locked-down browsers should also block access to locked screens where custom images can be shown

3.24 Locked-down browsers supporting remote online proctoring should detect and prevent technology threats to security.

(a) Locked-down, secure browsers may support remote proctoring by detecting virtual video, virtual microphones, and duplicate input devices.

Comments: Remote proctoring is a unique form of testing where the proctor is remote from the test taker. In addition to all the normal locked-down browser functionality, remote proctors typically need to block virtual cameras, virtual microphones, and machines with multiple input (keyboard/mice) devices. All these mechanisms can be used to fool remote proctors and cheat. Some tests need to access devices on the machine, such as a microphone, speakers, and/or video. While modern (HTML5+) browsers allow this to occur, the test taker will be prompted to give the application to access the device. Locked-down browsers can automatically enable this functionality for the test, eliminating the user prompt.

(b) Locked-down browsers should prevent a test from being delivered if external remote proctoring software ceases to run.

Comments: Remote proctoring software may run separately from the test being delivered in the locked-down browser. In these situations, the locked-down browser should have a method of validating that the remote proctoring software is running and be able to monitor the software so that the test is stopped if the remote proctoring software stops running.

3.25 Locked-down browsers should prevent test content from being stolen or exposed.

(a) Locked-down browsers should prevent screen captures of item content.

Comments: One approach to stealing content is to use software that records the screen while a test is being delivered. The first requirement for a locked-down browser to block screen recording is to use the built-in operating system functionality that blocks screen recording. This feature has the added benefit of blocking external remote-access software. Alternatively, for host systems that do not support this functionality, the locked-down browser should automatically block all unapproved applications from running and stop all processes that match known malware or screen capture software names. In addition to blocking screen captures, the locked-down browser shall support options to clear cut/paste buffers in memory before and after a test. In older operating systems, cut, copy, and paste functionality is limited to one fragment of content at a time. In newer operating systems, the user can cut multiple text fragments, and the operating system saves the text fragments in a memory queue that could be accessed after a test is complete. The locked-down browser ensures that any content copied during the test cannot be used after the test is over. The test delivery software is responsible for allowing or blocking the cut/paste functionality during a test.

(b) The locked-down browser should verify it is running an approved version of the software before delivering a test and automatically update itself before the test.

Comments: Installing software updates is intimidating for everyone involved. Programs and testing centers often want to avoid dealing with software updates to working software because changes inevitably lead to new problems. Unfortunately, security concerns override this desire. The locked-down browser should automatically update software and configuration test delivery. In today's rapidly changing threat environment, locked-down browser configuration should be continuously updated. Configuration typically includes process names or signatures that should be blocked or allowed.

(c) The locked-down browser should support configuration to clear cache before and/or after testing.

Comments: Test delivery systems will typically not cache content. Secure browsers can provide a second line of defense by automatically clearing cache from the domain and subdomains of the delivery system upon entry and exit of the test.

(d) The locked-down browser should provide configuration options to block proxy server attacks.

Comments: Normally, using HTTPS keeps content safe in transmission between the browser and the host servers. A proxy server attack is a method of intercepting HTTPS content. Such attacks require the test delivery computer to be modified to add a fake "certifying authority" certificate. In a proxy server attack, all content may be copied by the proxy server. The locked-down browser is the only defense against this attack. It can prevent this problem by (a) blocking all certificates that are expired or raise a security error (e.g., certificate name not matching the actual domain the content is coming from) and (b) verifying the certificates returned are using the public encryption keys expected from the test delivery servers by retrieving the keys separately as part of the secure browser configuration (i.e., "certificate pinning"). One challenge test delivery programs face is that many firewalls provide configuration options to use this technique to aid in virus detection, "sniffing" packets as they are returned to the browser. To prevent false positives, programs either need to turn off the secure browser check or require all test delivery locations to "whitelist" the test delivery domains in the firewall.

(e) Locked-down browsers should support the option to block assistive technologies (ATs) not related to designated accessibility features.

Comments: Most modern browsers support assistive functionality when test takers input text into a field. The locked-down browser functionality should be configurable, as some ATs, such as spellcheck, may be desirable, whereas others are not. However, the locked-down browser should not prevent the use of, nor affect the functionality of, approved accessibility assistive software.

(f) Locked-down browsers should block gestures that allow test takers to launch applications and access operating system features.

Comments: Gestures are movements made with smart touch devices such as screens or advanced mice. While useful shortcuts for users, gestures are a security hole locked-down browsers are responsible for blocking.

(g) Locked-down browsers should prevent unauthorized printing.

(h) The locked-down secure browser should be able to upload all actual or possible security violations detected to a central server.

Comments: This process will provide two functions. First, the data logged provides detailed information on the test event, including any possible violations. Second, the data allows the production teams to monitor unexpected or unknown scenarios and correct false positives that cause support issues.

3.26 Locked-down browsers should give test takers at testing locations a test experience that is compatible with most environments, prevents interruptions, and minimizes impact on privacy.

(a) **Locked-down browsers should support multiple operating systems, depending on the needs of the testing program.**

Comments: Education environments should support products such as Microsoft Windows, Google Chromebooks, Apple Macs, Apple iPads, and Android Tablets. Each of these platforms has unique challenges. Fortunately, some operating systems are beginning to include secure browser capabilities, led by mobile devices; however, this support is inconsistent.

(b) **Locked-down browsers should block automatic updates in the host operating system.**

Comments: Current operating systems typically run in an "evergreen" mode, meaning that they may be updated automatically at any time. Locked-down browsers should block all upgrades during test delivery to prevent the test taker from being interrupted.

(c) **The locked-down browser shall support uninstalling itself from the host operating system.**

Comments: For privacy reasons, it is vital to ensure the locked-down browser supports uninstalling itself from the host operating system so that it can be promptly removed from the test taker's machine following the test administration.

(d) **The locked-down browser should only be active while a test is being delivered, and all personal information that is tracked is limited to the test delivery. The information captured by the locked-down browser must be disclosed and documented.**

Comments: It is important to provide test takers with accurate details about any PI captured. In general, the locked-down browser should only track PI reasonably needed to fulfill its function. The locked-down browser should be less intrusive and limited to detecting processes and usage of hardware devices that pose risks to test content theft or cheating.

(e) **The locked-down browser should support different accommodation and accessibility needs.**

Comments: See also Chapter 10 (Fairness and Accessibility).

Guidelines for Interoperability

3.27 The user interface for test delivery should be designed to respond to the types of devices on which the test is intended to be administered.

Comments: If Smartphones, tablets, and the like will be allowed for test administration, the design of the user interface should accommodate such devices by displaying test items in a way most *appropriate for the screen real estate. Likewise, response interactions should allow for using the device's native features, such as two-finger pinch-zooming or touch screen interaction.*

3.28 TBA systems should either store data in an open, documented format or be able to export assessment data into an open, documented format that makes data available beyond the system's life.

Comments: Every system has an end-of-life point (e.g., technological obsolescence, commercial or practical reasons, a need to switch to a different system) that means it can no longer be maintained. At such a point, there will usually be a need to take data (e.g., questions, results) from the old system for use in a newer system or for reference purposes. Data should either be stored in a documented, open format or exported into such a format agreed to by the testing organization to avoid vendor lock-in or loss of data. Using an open format may also be helpful in complying with any obligations concerning data portability for PI under privacy laws.

3.29 Where industry standards or consensus specifications are available and suitable, those who control TBA systems should consider using them for data storage or export formats. This would make it easier to interoperate with other systems.

Comments: Using a standard or consensus specification makes it easier to move data from one system to another and reduces the chance of data misunderstanding or loss. It also reduces the risk that an organization may think it has a documented data format when not all the data are available or accessible.

3.30 Those who control TBA systems should consider using the IMS Question and Test Interoperability (QTI) specification as an export/import format for question data and, to a lesser extent, for other assessment data.

Comments: The IMS QTI specification has been in place since its first public release in 2000. It is a mature and widely used specification to import and export questions. To a lesser extent, it aids interoperability with other assessment data such as results (responses & scores) and usage data (item statistics). There are many versions of IMS QTI and different interpretations, but it is widely used for interoperability. For question and test data, as well as IMS QTI, consideration may also be given to the use of Moodle XML and for technology-enhanced items IMS Portable Custom Interaction (PCI), H5P (HTML 5 Package), and for CAT, the IMS Standard on CAT. For packaging, IMS Content Packaging can be used. The test items should be stored in formats compatible with and able to be repackaged with the maximum number of systems. For metadata, consideration can be given to IEEE Learning Object Metadata (LOM), Common Education Data Standards (CEDS), SIF (Schools Interoperability Framework) Data Model, DCMI Learning Resource Metadata Initiative (LRMI), and IMS Competencies & Academic Standards Exchange (CASE). Exchange of organizational and student data can leverage open standards such as IMS OneRoster, Ed-Fi Data Standard, and other local standards. Information determining the personal needs & preference for an assessment session can be exchanged using IMS AccessforAll (AfA) Personal Needs & Preferences (PNP).

Assessment content can be decorated with information to address special needs using IMS APIP and IMS QTI 3. Data related to digital credentials and pathways can be exchanged by use of open standards or specifications, e.g., IMS Open Badges and Open Pathways, Comprehensive Learner Record (CLR), CTDL (Credential Transparency Description Language), ASN (Achievement Standards Network), and Europass Digital Credentials Interoperability (EDCI). There are many other specifications and standards and work in progress to develop others, so this list is not exhaustive.

3.31 Where one TBA system or module/component calls another for delivery or scoring of an assessment, the integration method or API should be well documented.

Comments: Documentation means when one system needs to be updated or replaced, it is more likely the quality of the integration will be maintained. When planning and documenting the integration, consider error handling (e.g., what happens if one system fails), protection against spoofing (so each system can be sure it is calling or being called by the correct system and not an impostor seeking to subvert the assessment process). Also, ensure that neither system can be impacted by an "injection" or similar attack where computer code is used to try to disrupt a process--and that the called system maintains privacy required by the calling system. Other integration issues to be documented include scalability and potential overload if there are many simultaneous requests, the use of characters from different languages and special (e.g., punctuation) characters within text, and an audit trail of the call to allow troubleshooting and to show legal defensibility. What is required for each integration will vary depending on the integration use case, but the above is a useful checklist many integrations will need.

3.32 Where industry standards or consensus specifications are available, they should be considered when one TBA system calls another.

Comments: Industry standards are more likely to allow reliable integration over the longer term. If both systems support a standard, it is more likely that such support will be sustained over time. Such standards also allow easier substitution of other systems if required over time, whereas proprietary methods make it harder to switch systems. Industry standards will often but not always be more robust and secure than proprietary methods. Consensus standards or specifications to consider include IMS LTI, which allows a variety of calls from one system to another; AICC HACP, which allows launch and track of an assessment from a learning or other management system; ADL xAPI (Experience API) Standard, which allows tracking of assessment data and communication to another system; ADL SCORM (1.2 or 2004), which allows launch and track of an assessment from a learning or other management system; and HR Open Standards, which allows assessment interchange, particularly for recruitment systems. These five specifications or standards are widely used at this time; however, other specifications and standards are used in different contexts, and there is other developing work in this area. Others to be considered include IMS Caliper and ADL CMI-5. When using any standard or consensus specification that has conformance tools or a way of an implementation being certified, organizations should seek to have their implementation checked *against such tools and/or certified. This check will make the standard or consensus specification more likely to be implemented correctly. This recommendation also applies to data interoperability.*

Guidelines Addressing Test Disruptions

These guidelines are related to technology disruptions to tests. See also Chapter 8 for issues relating to test security and security incidents/data breaches.

3.33 Testing organizations and vendors should engage in comprehensive preventive activities designed to minimize the likelihood of any test disruptions during TBA administrations.

Comments: When developing systems, testing organizations and vendors should operate under the assumption technology disruptions in testing will eventually happen. This assumption should propel testing organizations to create infrastructure and systems to facilitate any reviews necessary when disruptions occur. When designing an assessment, testing organizations and vendors should identify the risks for disruptions associated with their program. For all risks identified, the testing organization and vendors should include well-defined activities underway or pending to mitigate against each of these risks. They should also develop systems for administering assessments that minimize the likelihood of testing disruptions, whether through system-wide failure or through the malicious activities of stakeholders attempting to disrupt the test administration. It is impossible to cover all aspects in these Guidelines, and thus readers are encouraged to review more comprehensive resources such as CCSSO and ATP (2013), ISO (2019), ITC (2005), Luecht (2015), and Martineau and Dadey (2016). Pilot testing is encouraged to consider test disruption risks and ensure appropriate data and information are collected to evaluate the degree of impact of each risk. In addition, testing organizations and vendors should analyze all resulting data collected from the pilot, including identifying any locations with potential connectivity or compatibility issues.

3.34 The testing organization should develop a response plan in the event that a TBA disruption occur.

Comments: The response plan for technical disruptions should include information such as the roles and responsibilities of people involved, including initial and final decision-makers, individuals who fill communication roles, and those who will need to be kept informed throughout the process. The response plan should be consistent with the purpose and use of the assessment and the specific data and information used in any validation argument for the testing program. Organizations can further test response plans by simulating disruptions and responses. Testing organizations and vendors should coordinate in how to identify disruptions and what steps should be taken by those responsible for any immediate and time-sensitive decisions as the disruption is identified, as well as designate the individuals responsible for sign-off for any final decisions regarding the impact of any testing disruption. It may be helpful to have templates for communication already developed and documentation for how all relevant data and information can be compiled. Note: Information on security incident response plans is found in Chapter 8.

3.35 Testing organizations and vendors should have clear policies regarding who is authorized to provide communication for the organization when a disruption occurs and how this information will be transmitted.

Comments: When considering policies for communication, organizations should acknowledge any time a testing disruption occurs, various degrees of information will be shared with the general public through social media and other means at a rapid rate. The communication policy must allow the testing organization to provide the status of any testing disruption quickly and clearly, especially any response taken that may impact test takers.

3.36 While developing TBA administration systems, testing organizations and vendors should build systems for the proactive gathering of data that could identify any testing disruptions in real time during any test administration.

Comments: These systems can detect data such as bandwidth capacity, Internet connectivity, and other similar data points. These systems should also have procedures in place to notify the appropriate individuals if the data indicate testing disruptions may be happening. When developing systems for the proactive collection of data and information during test administration activities, testing organizations and vendors should also develop a clear set of procedures to be followed throughout the entire test administration process. These activities should include the data to be collected, the procedures for monitoring all data, and the criteria for flagging any indicators for potential testing disruptions. When developing systems for proactively collecting data and information, procedures should be developed for defining what activities should be followed if any data indicate an incident may have occurred. These procedures can include contacting testing centers for additional information, additional data analyses, or reaching out to test takers for further information.

3.37 Individuals charged with monitoring and administering a TBA should be provided comprehensive training on all aspects of the test administration, including the steps to follow in the event of a test disruption or any other critical irregularity.

Comments: Individuals charged with serving as support for a testing program should be provided clear directions for how requests for assistance should be documented and resolved. These directions should specify roles and responsibilities for escalating irregularities to senior management.

3.38 In the event of any testing disruption, testing organizations (with the assistance of testing vendors, if any, and optionally, a qualified independent party) should conduct a comprehensive, independent investigation of the impact of the testing disruption.

Comments: When feasible and appropriate, testing organizations and testing vendors may use the services of an independent party to assist in conducting the investigation. The independent party *could complete the investigation activities or serve as a reviewer of the activities completed by the testing organization.*

3.39 In the event of a TBA disruption, a testing organization should refer to well-developed purpose statements for its tests because these statements will provide essential guidance for the evidence needed to determine the impact of the testing disruption on the use of test scores.

3.40 A testing organization should ensure data can be made readily available to facilitate the review and remediation of any testing disruption investigation.

Comments: Information regarding the location and format of all data, the methods required to extract and share the data, and the completion of all appropriate documentation should be readily available.

3.41 A testing organization should proactively incorporate planning for technical disruptions in testing when entering into a contractual relationship with a vendor, when making changes to administration plans, or when releasing a request for proposal for TBA services.

Comments: When testing organizations release a request for proposal for TBA-related services, they should require all respondents to provide detailed plans for how they will work to prevent technical disruptions and detailed plans for activities in the event of any testing disruption. When vendors develop systems, policies, procedures, and/or other resources relevant to addressing testing disruptions, they should clearly document consistency with these guidelines. That documentation should be provided either as a response to request elements addressing testing disruptions or as a part of a base package, a value-added service, or an additional cost option.

3.42 Testing organizations should be transparent with their stakeholders and the general public when a TBA disruption occurs.

Comments: It is important for testing vendors to be as transparent as possible with the testing organization they serve. When communicating with the public, testing organizations should not attempt to minimize or downplay the impact of any testing disruptions, especially when information about the nature and scope of the disruption is still being collected (likewise for testing vendors, when communicating with the testing organization). When communicating with the people immediately impacted by any testing disruption, testing agencies should acknowledge the disruption that led to difficulties for test takers and potential test score users. However, no specific plans for any remediation to test takers should be discussed or speculated until a more thorough investigation is completed. In some cases, it may be helpful to create categories or groups if different test takers experienced different types of disruptions. In such cases, the specific issues may be communicated to the public in a manner that recognizes these different experiences. It is likely or at least possible test vendors and test agencies will seek advice from legal counsel on the occurrence of such events.

3.43 In some cases of TBA disruptions, testing organizations should carefully consider the possibility of providing another testing opportunity to be chosen by the test taker or stakeholder leadership.

Comments: In instances where test takers need to pass the examination to practice or gain employment, the testing organization should develop a plan to allow impacted test takers a prompt opportunity to retest. Fee waivers for retests are likely to be appropriate in many instances.

3.44 A testing organization should develop clear criteria that define whether any testing disruption has resulted in a significant loss of fidelity for its testing program.

Comments: The nature and scope of the testing disruption should be explicitly connected to the validity arguments the organization has developed for using their test scores.

4. SCORING

Background

Advances in technology have enabled new and enhanced capabilities in automated scoring of assessments, including scoring of selected responses, technology-assisted human scoring of constructed responses, and fully automated scoring of constructed-response assessments. In addition to the content of test taker responses, technology-enabled modeling of response time has also proven useful in improving estimates of test item and test taker characteristics (e.g., difficulty, ability). These technology-enabled scoring capabilities require sound and dependable systems. Thus, it is important to consider the possibility of disruption and incomplete scores and take appropriate steps to identify and address such cases. This chapter extends existing standards for scoring assessments (AERA, APA & NCME, 2014, Ch. 4; Educational Testing Service, 2021), emphasizing considerations for technology-based assessment (TBA).

Automated Scoring of Selected Responses

Selected-response (SR) items are assessment items where the test taker is asked to choose an answer from a finite set of options. Traditional SR items include the true-false and multiple-choice items, where the test taker selects one or more responses out of a larger number of provided responses. Test developers have created new types of SR items that extend beyond the traditional true-false or multiple-choice item (see Chapter 1 for further details regarding technology-enhanced item types).

The automatic scoring of SR items on assessments traces back to scoring multiple-choice items via mechanical scanners starting in the 1930s and optical scanners starting in the 1960s. Today, computer-automated scoring engines can handle traditional and computer-based SR items quickly and effectively. *Automated scoring* utilizes computer technology to apply customized rules or statistical models that are designed to input test takers' responses and output scores that emulate what a human scorer would assign based on the scoring rubric. Ideally, scores returned from an automated scoring program should be indistinguishable from the scores assigned by a human scorer. To this end, automated scoring should be based on clear, vetted rules that match the rubric's expectations. For many item types, this would include identifying all response combinations that yield full credit or partial credit, and the scores to assign in each case. In some engines, feedback may also be provided through a set of codes. Automated scoring rules for an item should achieve 100% agreement with an independent human rater whenever possible.

By writing unambiguous item rubrics and automated scoring rules for SR items on computer-based assessments, assessment stakeholders can use any SR item format while assuring scoring can be done automatically, reliably, quickly, and with minimal error.

Automated Scoring of Constructed Responses

This section describes the use of advanced methods for automated scoring of constructed-response (CR) items using computer algorithms and rubrics to derive scores from unconstrained, open-ended test item responses. The goal of automated scoring is typically to emulate human scoring. Automated scoring can be applied to a range of item types and input modes. Item types include short and long CR items, including numerical responses, cloze tests, and essays. Item input modes include text and speech. Automated scoring in assessment programs is motivated by practical needs, including cost reduction, faster score return, and the mitigation of rater staffing shortages. Automated scoring can also address measurement and fairness considerations by helping to ensure scoring consistency within and across test administration windows. The following summary describes a common approach to automated scoring of CR in writing assessment.

Design. Automated scoring systems have typically been designed using a three-stage process, and each process should be consistent with the rubric criteria. The first stage involves normalizing test takers' responses to better identify relevant linguistic or structural segments, such as characters, words, sentences, and paragraphs in the case of text, or to identify phonemes or sounds more accurately in the case of speech. This normalization process may involve text cleaning, such as removing extra characters, replacing characters or tags with other characters recognizable by the system, and correcting misspelled words. The text cleaning process should be consistent with the rubric criteria. For example, spell correction may be appropriate only for items for which correct spelling is not germane to the scoring rubric criteria being assessed; it is not appropriate when the quality of spelling is relevant to the rubric.

The second stage involves extracting features from the normalized text that reflect the relevant rubric criteria. Feature development can be theoretically driven whereby features theorized as important are developed using computational methods; alternatively, machine learning methods can be utilized to "learn" relevant features. In the case of writing characteristics, features may consider grammar and mechanics, vocabulary usage, discourse phrasing, word choice, cohesiveness, and organizational elements. In the case of content, features may be the patterns of words or phrases associated with rubric criteria levels. For speech scoring, some features may overlap with writing characteristics (such as vocabulary usage), but speech-specific features may include rate of speech, elements of pronunciation, and fluency. The third stage involves training a statistical model to best predict human scoring using the features computed in the second stage.

Development. When automated scoring systems are developed or "trained" on samples of responses it is critical that the samples are representative of important characteristics of the population to be tested and that there is rigorous training and monitoring of the human raters who will ultimately complete the score assignment. At a minimum, responses should be scored by two independent raters to allow for a comparison between the automated scoring/human performance and human/human performance. Samples used to train and evaluate automated scoring systems are typically in the thousands of responses per item. However, this number may vary based upon the score point distribution and the number of examinees involved in a testing program. These samples are typically divided into training, test, and validation subsamples. Multiple competing models may be built using the training sample and evaluated on a test sample; the best-performing model is selected based on test sample performance. The best-performing model then scores the validation sample, and those scores are used to evaluate automated scoring performance, typically by comparing to human scores. In more complicated designs, multiple models are built in parallel, and scores from each model are combined statistically to produce an ensemble of scores. The ensemble's performance is then examined using the validation sample that is representative of individuals to be tested.

Performance. Automated scoring performance is typically evaluated by examining the extent of agreement between the automated system-generated scores and the human scores, with the goal of reproducing human scoring. Thus, systems should have similar agreement and score distributions to those produced by humans. Common measures of agreement include exact agreement, quadratic weighted kappa, and root mean square error standardized mean difference a common measure of score distribution similarity (Williamson, Xi, & Breyer, 2012; Yannakoudakis & Cummins, 2015). It is recommended practice to involve program stakeholders (e.g., clients, psychometricians) early in the automated scoring process when defining the performance evaluation measures and criteria, while considering their utility.

Typically, the criteria for agreement between human and automated scoring are defined by *relative* or *absolute* thresholds. A *relative threshold* would specify, for example, that the exact agreement of the engine with a human rater should be no more than 5% lower than the exact agreement of two human raters with each other. Another example would be that 90% of human and automated scores are within one score unit. An *absolute threshold* would specify, say, that the exact agreement of the engine with a human rater should be no lower than 70%. A core psychometric principle is fairness. The performance of automated scoring systems should be evaluated for test taker subgroup populations to ensure scoring consistency holds across groups of test takers defined by gender, race/ethnicity, and other personal characteristics, including individuals with disabilities. (See also Chapter 10. Fairness and Accessibility).

In addition to predicting scores, automated scoring systems should be able to identify unusual responses. Unusual responses can take the form of non-attempts, particularly creative responses, responses due to a disability and use of assistive technology, or bad-faith responses such as those written to try to trick the system into producing a higher score. In speech scoring, unusual responses could be poorly recorded, uninterpretable, or submitted by multiple speakers. Other unusual responses can be plagiarized or disturbing responses where intervention is recommended to protect a test taker's safety. Custom filters are used to identify such responses. These filters play a critical role in ensuring test takers receive fair scores when a follow-up plan is developed and implemented.

Automated scoring systems can be used with or without human scoring during live test administrations. It is recommended to include some amount of human scoring alongside automated scoring for quality control (QC). The inclusion of human scoring supports the ability to monitor the automated scoring quality in the event of a technical error, rater drift, or change in test taker population or test taking behavior. Equally important, certain types of unusual responses may be better scored by humans than by automated scoring; identifying and routing such responses will improve overall score quality. In addition, the inclusion of humans in the decision-making loop is an essential component of AI systems (see Part IV).

Programs should regularly evaluate automated scoring performance to ensure continued scoring quality and provide mechanisms for program stakeholders to ask questions about the scoring process. Technical documentation around the performance of automated scoring can support this monitoring.

Technology-Assisted Human Scoring

Human scoring of CR items has evolved as approaches have shifted from face-to-face, on-site models to distributed online systems. This is true of traditional assessment formats such as written tests of academic content and more complex performance tasks traditionally requiring live observation. Hand-written essays and worked solutions to mathematics problems, once marked in face-to-face sessions, often are now digitized and presented to raters online. Other assessment output can be presented to human scorers in their original format within an online scoring software platform, such as keyed essay responses, digital files of spoken item responses, computer code sets, or digital images of artwork. Even very complex performances, skills, and processes traditionally observed *in situ* are more frequently being captured in video, simulation, or virtual reality contexts for asynchronous evaluation. These include such things as music, dance, or theater recitals; gymnastics routines; field-sports activities; classroom teaching practices; surgery; and customer-service interactions. Such complex operations may still be assessed live during the act by a trained observer where alternative capture formats are not considered sufficient.

Online human scoring may be completed on-site or remotely distributed. The same is true of rater training and ongoing calibration, resulting in some mixed-mode circumstances where raters are trained and score on-site, are trained on-site and score remotely, or are trained and score remotely. Raters, like many learners, tend to prefer on-site training and scoring (Hamilton, Reddel, & Spratt, 2001; Kunin, Julliard, & Rodriguez, 2014), but the two modes generally deliver results of similar psychometric quality (Kemp & Grieve, 2014).

Systems. Online scoring relies on access to a software platform for data collection. This system should be supported on more than one common browser and operating system. Failure to use a well-designed display and data collection platform is likely to result in more data-entry errors, complex manual scheduling procedures, and rater/response assignment issues, as well as longer timelines for scoring. Full-featured scoring platforms offer a range of tools to manage human scoring and typically include a means for ongoing human scoring calibration through the delivery of seeded exemplar exams.

For online scoring, technological equipment is essential. There may be minimum requirements for computer equipment used in remote marking, such as monitor size or resolution, speed of home Internet connection, and capacity to complete software installation and system sufficiency checks. These skills should be made clear as a condition of hiring scorers. On the other side of the system, electronic file storage could become burdensome if the files submitted are large, there are numerous responses per test taker, and/or the assessment is given to a large number of test takers. File storage of CRs often has significant security constraints, and these vary substantially by location (see Chapter 6 for more details on data management).

Scoring in the Event of Technology Disruption and Incomplete Assessments

TBAs present new opportunities for test providers and users, but they also present some challenges. One of the most important challenges is overcoming technological disruptions such as server failure and other testing interruptions such as delayed login and involuntary logout, which often lead to incomplete testing sessions.

Incomplete testing sessions, especially those missing due to technological disruptions, lead to the problem of missing or incomplete data and pose a problem in the reporting of fair and valid scores. The testing organization has the option of not reporting a score to affected test takers, especially for those for whom the extent of incomplete data is severe. The extent and severity of incomplete data can be determined by one or more of several statistical or psychometric measures. These include the reliability of the score on the "complete" part of the test and the bias and standard error of the score that can potentially be assigned to the test taker based on the "complete" part of the test. Depending upon the purpose of the assessment and the extent of incomplete data, an imputation approach may be used in some cases to estimate scores on the missing parts of the test (in some ways similar to computer-adaptive testing). However, caution should be exercised in using this approach appropriately and ensuring the resulting imputed score is not biased or erroneous and is fair and valid. Imputation may be controversial and ill-advised in high-stakes decision-making where legal defensibility is a concern.

These guidelines do not address the common issue of omitted or not-reached items not caused by technological disruptions. These cases have been examined by researchers (De Ayala, Plake, & Impara, 2001; Finch, 2008) and are outside the scope of these guidelines. Consequently, "incomplete" henceforth will refer to "incomplete due to technological disruptions."

Using Item Response Time in Scoring

Response times can also be part of the scoring rule of the test when timing is an important attribute of performance (Maris & van der Maas, 2012; van Rijn & Ali, 2018). For example, Klinkenberg, Straatemeier, and van der Maas (2011) considered a score rule in which the points awarded to a correct response are higher for faster responses and lower for slower responses. In addition, points are subtracted if the response is quick and incorrect. Due to the resulting correlation between responses and response times, precision of the trait estimates can be improved (e.g., van der Linden, Klein Entink, & Fox, 2010; Bolsinova, & Tijmstra, 2018), although the benefits might be limited (Ranger, 2013). Once the test design is determined and the response times have been collected, a suitable statistical measurement model may be determined for the analysis and scoring. If the responses play a (key) role in the analysis (which is the case if the times are collected within the collateral information design, for example, but not if the response times are collected as the sole measure of the trait), the responses and response times should ideally be considered in a simultaneous modeling approach (e.g., van der Linden, 2007; Molenaar Tuerlinckx, & van der Maas, 2015). Within the measurement model of choice, response times (and the responses if applicable) can be analyzed and scored. Chapters 7 and 10 describe psychometric quality and fairness considerations that are especially important when contemplating the use of response time in scoring.

Guidelines for Scoring Technology-Based Assessments

Guidelines for Automated Scoring of Selected-Response Items

4.1 During item development, item writers should consider the scoring method to be used for the testing program.

Comments: For example, most dichotomously scored item options are written to have clearly correct and incorrect answers. If partial credit scoring methods are used, item options are constructed with clearly defined and varying score levels.

4.2 Item rubrics should be clearly defined for each item, including scores for all response options in SR items. If scores are aggregated across parts of a response to produce an item score, this should also be clearly defined in the rubric.

Comments: For polytomously scored items, it is advisable for the rubric to detail scores for all possible response patterns. Hotspot and Limited Figural Drawing items require that points or areas in the graphic stimuli representing the correct, partially correct, and incorrect responses be clearly defined in a format machine readable by the scoring engine. For the highlighting item type, all highlightable (clickable) words and phrases are tokenized in machine-readable format, and score values of each token are clearly defined. For arithmetic and calculation items, the correct answer(s) may include a range of acceptable responses that account for allowable rounding errors, where applicable. The display unit of measurement and the number of decimal places allowed for item responses in the item stem are important considerations. Additional considerations for equity and fairness are outlined in scoring Chapter 10.

4.3 Testing programs using automated scoring for SR items should establish policies that are published and shared with test takers and other stakeholders.

Comments: It is recommended a testing organization adopt scoring policies that cover areas such as revision of scoring rubrics or item retirement if an error is discovered or the item content has changed; rescoring individual items and the assessment as a whole if an error is discovered; and the test takers' right to review items or challenge scores if review and challenge are allowed by the testing program. To the extent that some of this information does not disclose proprietary intellectual property, and thus may be sharable with test takers, it should be included in the organization's Test Taker Agreement (or similar form).

4.4 QC procedures and rules should be conducted and documented before using automated scoring. QC responsibility should be shared among appropriate staff responsible for the assessment.

Comments: By having a team of professionals review the scoring rules before using the items in the field, any last-minute cases can be identified and corrected before operational use of the items. These include automated scoring software developers, machine learning scoring trainers, item writers, test developers, and psychometricians. *The following steps are recommended QC procedures:*

- *If an automated scoring program or AI-based scoring engine is introduced, compare its performance with the testing program's existing scoring method (e.g., a different machine scoring engine or scoring by human raters). Review interrater reliability between new and existing scoring methods.*
- *Subject matter experts review item scoring rubrics for content accuracy and adherence to the testing program's item referencing guidelines.*
- *Automated scoring software developers, engine trainers and psychometricians review whether the scoring rubrics are correctly applied in the software or machine scoring algorithms.*
- *Review score reports and other scoring outputs for accuracy.*
- *Review item data and metadata files to ensure all scoring data are correctly associated with the items.*
- *Conduct the same QC procedures if the scoring software or AI engine undergoes any updates or enhancements.*

4.5 After items and the automated scoring software or AI engine become operational, appropriate personnel should review a representative sample of test takers' responses and the scores assigned.

Comments: There should be one-to-one correspondence between a response and the score it receives for all groups of test takers; each selected response should receive one, and only one, score from the system.

4.6 During the operational administration, testing programs should establish a regular cadence to check live testing data for scoring accuracy. The cadence may vary depending on testing program policies, methods of test administrations, examinee volume, number of test forms, and other factors. The process and cadence for operational scoring data verification should be documented.

Comments: Possible item quality indicators to calculate include score point distributions from the automated scoring software or AI engine, which should be comparable to distributions observed in past administrations or field testing; item difficulty and discrimination statistics; agreement between scores assigned by the automated scoring software or AI engine and those by human raters; and kappa or quadratic weighted kappa between the automated scoring software or AI engine and human raters. It is important to calculate these quality indicators for subgroups of test *takers such as racial/ethnic groups, individuals with disabilities, and second language learners when feasible*.

4.7 **If an incorrect score is assigned, the issue should be escalated to appropriate personnel for resolution, and the scoring software or AI engine rules should be updated as soon as possible to prevent further incorrect scoring.**

Comments: Scoring rule updates may not be possible until assessment stakeholders have given permission to make the change.

4.8 **QC steps should be regularly repeated to verify the accuracy of the scoring rubrics and software or the AI scoring algorithms.**

Comments: Post-operational quality controls help ensure the content of the items continues to reflect current and accurate knowledge and confirm the scoring software or AI algorithms did not inadvertently change after they were programmed.

Guidelines for Automated Scoring of Constructed-Response Items

4.9 **The rationale for using automated scoring should be clearly articulated and appropriate for the program in which it is used. Documentation of the design and use of the automated scoring system should be developed so stakeholders can make reasonable decisions about the scope of its possible application and use.**

Comments: Rationales may include cost savings, faster scoring, rater staffing, accuracy, and reliability. Appropriateness evaluations may include how the engine design, performance, and methods for combining human and automated software/AI engine scoring support program goals and stakes. Documentation should be written so that the relevant program stakeholder groups (e.g., technical vs. non-technical audiences) can understand it.

4.10 **The automated scoring system design should appropriately reflect the constructs assessed via the items, rubrics, and other scoring materials. The relationship between the scoring system and the constructs assessed should be documented, including the flow of responses through the 3-stage process (normalization, feature extraction, statistical modeling), detection of unusual responses, and design limitations**.

Comments: Documentation normally includes how the process reflects the rubric criteria. It is recommended that unusual responses (e.g., non-attempts, creative responses, gaming responses, plagiarism, disturbing responses, and responses by speech impaired and second language learner test takers) be described, including how the detection of these types of responses fits into the overall architecture. Documentation may include design limitations (e.g., automated scoring software and some AI scoring systems do not understand language and so may behave in unexpected ways to bad-faith or unusual responses).

4.11 If an AI scoring system requires training, it should be trained on a representative sample of responses that have been human scored with the highest level of quality the program supports. The rating process should be monitored and aligned with the program's operational practices, with clear measurement and construct validity criteria used to evaluate the quality of the scoring.

Comments: The sample selection should consider the operational context, size relative to training and validation needs, and appropriate representation of program-identified key subgroups to ensure diversity and avoid potential bias. The methods and materials used to hire, train, qualify, monitor, and retrain raters reflect the best practices used in the program and adhere to agreed-upon clear performance criteria. A minimum of two raters is recommended for each response in the training sample to support the evaluation of automated scoring performance relative to human scoring performance. Consider compliance with privacy laws when using test taker data for training data sets.

4.12 The validity, reliability, and fairness of automated scoring should be evaluated using sound methodological and statistical approaches and clear evaluation criteria. The methods and procedures should be documented and provide recommendations about appropriate use of the automated scoring system.

Comments: Following are recommendations for ensuring valid, reliable, and fair automated scoring of CR items:

- *Establish procedures to develop and validate the automated scoring system. Consider the inclusion or exclusion of aberrant responses, the methods for creating development and validation samples, the choice of score used as the dependent variable, the rationale and use of various trained models, and the rationale underlying final model selection. If an AI system is used for automated scoring, the testing organization should also develop a method for training the scoring engine to make sure it is valid and fair.*
- *Evaluate the performance of the automated scoring system on a validation sample that represents the test-taking population. In the case of AI scoring, the validation sample should be independent of the training set used to build the scoring model. Using the validation set, examine the level of agreement with the human raters and compare this agreement to the level of agreement between a minimum of two human raters.*
- *If an item is assigned multiple scores, examine all score relationships and patterns within the item.*
- *Examine relationships between scoring program/engine features and non-construct-relevant features, especially when such features may be commonly acknowledged as having a strong relationship to the item score.*
- *Evaluate the measurement accuracy of unusual response identification methods.*
- *Examine the performance of the automated scoring system for program-identified groups, with considerations for instabilities around small sample sizes and ability differences.*

4.13 The approach for using automated scoring or human scoring methods during test administrations should be based upon the scoring performance of each method and consistent with the goals and stakes of the program. Describe the rationale for the approach and the methods for combining automated scoring or human scoring.

Comments: Approaches can include fully automated scoring with no human review, fully automated scoring with some human review, partially automated scoring with some responses routed for human scoring, combined automated and human scoring whereby each response receives a score from both sources, and fully human scoring. Approaches may be configurable at the item level to account for the fact that an automated scoring system may perform well for some items and perform poorly for others. Item-level configurations might include whether to use automated scoring or thresholds for when responses can be auto scored or human scored. Suppose a change in approach occurs (e.g., fully human scoring to fully automated scoring) within an already-defined program. In that case, it is important to compare the new approach to the original scoring approach to investigate, identify, mitigate, and document any potential impacts on test scores, item parameter estimates, and achievement levels.

4.14 A well-defined process for reviewing automated scoring performance during and after test administrations should be developed, documented, and implemented. There should be a process in place for handling errors or disruptions.

Comments: The engine monitoring process often includes evaluations of agreement with human scores and the proportion of responses routed for human scoring. The capability to answer questions from stakeholders about scoring is an important consideration. It is recommended the automated scoring system be able to provide data that show how the system arrived at a score for a given response. A written mitigation plan is recommended for automated scoring errors due to unexpected infrastructure or scoring issues.

4.15 For AI-based scoring systems, algorithmic predictions, recommendations, outcomes, and prescriptive actions should be derived via transparent, ethical, and bias-free methods that can be explained and evaluated by internal and external experts or expert systems.

Comments: "Right to understand" and "right to forget" are important considerations for technology-based assessing organizations. Approaches such as InterpretML and other transparency systems allow for inspection and attempt to enable explanation of outcomes from otherwise opaque prediction models. While expert systems and related implementations can be reconstructed based on nodal data removal, neural network or other associated weight-based models can be quite difficult to retrain absent a data set that has been "forgotten." Right to forget can be very difficult in learning model or graph/node model systems, given the interconnections between data elements.

Guidelines for Technology-Assisted Human Scoring

4.16 Design of human scoring processes should consider key design factors, such as rater qualifications, scheduling, and work structures; assuring anonymity of the test taker; rater accuracy and consistency checks; QC; and scoring of multiple item submissions.

4.17 Rater training should be thorough, including using the scoring system, avoiding potential biasing factors, and evaluating training effectiveness through assessment of rater accuracy.

Comments: Human raters may have particular tendencies such as overuse of central or extreme score categories or highly variable scoring. These are generally reduced through effective rater training. CR scores, especially essays, may be influenced by numerous factors: response mode (keyed versus hand-written), writing style, length, use of complex vocabulary, grammar and typographical errors (when not part of the construct being measured), and candidate choice of topic (if allowed). Scores for image, audio, or video responses may be affected by recording quality, lighting and sound levels, and candidate appearance, among other factors. Additional potential biasing factors include rater knowledge of the test taker and knowledge of previous scores assigned, as well as the rate of scoring resulting in a disproportionate number of responses scored, and fatigue that my result from scoring for an extended period of time.

4.18 Technology platforms used for response display and score capture should be user-friendly and include the scoring assignment and management tools needed to facilitate high-quality rating results.

Comments: It is important for the platform to provide raters with a complete set of resources to apply the scoring criteria. The platform should enable raters to make comments without changing the test takers' responses. The platform may also provide rater metrics in a dashboard that may be used to monitor rater performance.

4.19 Technology requirements for participation in scoring should be clearly defined and made available to potential raters as part of recruitment.

Comments: There may be minimum requirements for computer equipment used in remote marking, such as monitor size or resolution, speed of home Internet connection, and capacity to complete software installation and system sufficiency checks. It is helpful to make these requirements clear as a condition for hiring scorers. On the other side of the system, electronic file storage may become burdensome if the files submitted are large, there are numerous responses per candidate, and/or the assessment is given to a large number of candidates.

Guidelines for Scoring in the Event of Technology Disruption and Incomplete Testing Sessions

4.20 Appropriate equipment, processes, and procedures should be in place to prevent technological problems (disruptions) and to minimize the extent of the adverse impact of such disruptions for TBAs.

Comments: Martineau and Dadey (2016) provide an excellent list of recommendations on how to adhere to this guideline. One example of appropriate equipment is independently operating databases on independent servers that exist for the sole purpose of documenting issues with test administration. It is important to include policy/procedure for test takers to contest scores or pass/fail decisions (see Chapters 8 and 9 for additional guidelines related to test taker appeals).

4.21 Data collection systems used for tests should be able to document all technological disruptions, including information about testing interruptions for each test taker.

Comments: The extent of loss of data will often depend on the extent of documentation.

4.22 If technological disruptions occur, all attempts should be made to recover data to minimize the extent of loss of data.

Comments: For example, it is recommended that data from all servers be combined in case the main server does not include all item response data.

4.23 If technological disruptions result in incomplete data for some test takers, the testing organization may, in certain cases, use an approach for imputation/projection/estimation of the missing scores. However, such approaches should be validated by empirical research.

Comments: If scores are reported to test takers with incomplete testing sessions, the assessing organization should ensure the procedure to impute and report the scores is rigorous and produces scores that are reliable, valid, and fair--and is appropriate for the purpose and stakes of the assessment. In some cases, imputing scores may be defensible based on the results of a thorough comparison. Imputed scores should demonstrate validity (i.e., same scores are produced for examinees who had complete data when their data are deleted and imputed for comparison) and reliability (the reliability and an associated standard error of measurement can be computed using the "complete" part of the test). The validity of the imputed scores can be established by confirming that the content coverage of the "complete" part is like that of the total test and by showing that the correlation between the imputed score and an external criterion is very close to what the correlation between the total score and the external criterion would have been without any missing scores. The fairness of the imputed scores can be established by, for example, showing the procedure used to impute the missing score does not introduce any bias in the reported score and that relevant demographic groups are not disadvantaged by the imputation procedure.

4.24 If technological disruptions result in incomplete data for some test takers, the testing organization may choose not to report any scores for those test takers. Before taking the decision of reporting no scores, the testing organization should evaluate whether the decision would have any negative consequences on test takers or other users of the test scores and try to mitigate any possible negative consequences.

Comments: When a technological disruption results in incomplete data, it is common for testing agencies to allow a free retest when it is convenient for test takers to retest.

Guidelines for Using Item Response Time in Scoring

4.25 If response times are used in scoring the test, this should be disclosed to test takers and others who interpret test results.

Comments: It is important to consider communicating factors that affect scoring (e.g., accuracy, testing time, wrong answers) to avoid differences between groups in the degree to which they understand the scoring factors and ensure that an emphasis on response times does not cause undue stress to certain subgroups (e.g., subjects diagnosed with dyslexia, non-native speakers, older age groups, etc.). It is also important to ensure test takers are not penalized due to latencies in data transfer due to platform or Internet delays.

4.26 Recording of response times should be as accurate as possible, avoiding the effects of technology requirements to respond to an item, testing in an environment free of distraction, and measuring time with a high degree of precision.

Comments: It is important to avoid individual differences in response times due to construct-irrelevant differences in computer technology and distracting factors in the testing environment that may distort the response process. In the measurement of response time, precision is important (e.g., milliseconds are preferred over seconds).

4.27 Construct-irrelevant factors that may affect response time, such as motor disabilities, executive function challenges, testing in non-dominant language, and other personal characteristics, should not negatively impact test takers' scores.

Comments: The logic of using item response time in scoring is to measure processing speed when it is construct relevant. Construct-irrelevant factors may also influence the time it takes time to respond to items, and test takers should not be penalized for these factors.

4.28 Acceptable fit of the response time model should be established before using the model in scoring, including the appropriateness of the fit of the model across groups. Detection and removal of outliers should also be considered in evaluating model fit.

Comments: Model fit is an important consideration before using a response time model, including appropriateness across groups. Differential item functioning can be used to determine if the response time model is appropriate across groups of test takers potentially affected differently by the response time instruction (e.g., respondents diagnosed with dyslexia, non-native speakers, older age groups). Response-time outliers may be removed if clearly due to technical failures.

5. DIGITALLY BASED RESULTS REPORTING

Background

Testing organizations have a responsibility to report accurate scores to individuals and organizations. Digital reporting of test results is increasing and is state of the art in technology-based assessment ("TBA"). This electronic communication of test results dramatically shifted the narrative of traditional paper-based "score reporting." Results reporting commonly takes the format of a query uniquely devised by a user of the data, particularly in the context of group-level reporting. In this dynamic (interactive) reporting approach, users engage with varyingly sophisticated online data analysis tools and large data repositories to create their own queries and answer their own data questions (Zenisky & Hambleton, 2013). Moreover, the timing of score reporting for TBAs is often immediate, so a test taker can obtain a score once the test event has ended.

Neither static nor interactive reporting should be viewed as secondary or inferior to the other approach. Rather, both should be viewed as complementary strategies that are appropriate for different audiences and different uses of assessment results (uses that are themselves psychometrically valid and appropriate).

These guidelines for digitally based reporting are informed by the Hambleton and Zenisky model of score report development (Hambleton & Zenisky, 2018; Zenisky & Hambleton, 2015), where the basic principles of purposeful development in reporting apply. However, the guidelines are specific to TBAs where results are reported digitally.

Maintaining Confidentiality of Score Reporting

Individual test scores are typically confidential and need to be treated securely. Aggregate scores (e.g., averages across assessment participants) are typically not personal data but may still be confidential to the test sponsor and, if not properly aggregated, can inadvertently reveal personal information. Consistent with the AERA et al. (2014) *Standards* and data privacy laws, organizations must maintain data security protocols to protect confidentiality. The guidelines in this section aim to establish best practices in maintaining the confidentiality of score reporting for TBAs. Please also refer to Chapters 8. Test Security, and 9. Data Privacy.[4]

[4] In some instances, the confidentiality of a test taker's scores is not the decision of the individual but, rather, the customer of the testing organization (e.g., employer) may decide on when and how to share the scores. In other situations, the test taker may give the testing organization permission to share scores with others (e.g., certification bodies, potential employers).

Guidelines for Digitally Based Results Reporting

Guidelines for Results Reporting

5.1 Data quality procedures should be established to ensure results transmitted to test takers or other stakeholders at the conclusion of a TBA are accurate.

5.2 Policies and procedures should clearly define the different types of reported scores such as raw scores, scaled scores, performance classifications (e.g., pass/fail), and other individual scores derived directly from the TBA.

5.3 Digital reports, when printed or exported, should be date stamped, identify any filters used, and indicate sample sizes, where applicable.

Comments: These details help users understand when the report was created and may assist with auditing discrepancies between different versions of a similar report.

5.4 The business rules governing when to include or exclude data from digital or static reports should be documented and available for review.

Comments: If tests completed in too short a time are not included in the reporting database being queried, that decision rule should be documented so users are aware of the rules used to include or exclude data.

5.5. Consideration should be given to the test purpose, audience for results, and test use, to integrate existing reporting guidelines and research in the development of reporting materials.

Comments: Different report formats (summary, highlights, full-length reports) and data displays (text, graph, and tables) should be considered to ensure materials are understood and the conclusions being drawn are appropriate. Using focus groups to create better reports is recommended.

5.6 Report documents should be piloted with stakeholders using a variety of data collection techniques for accessibility, usability, and understanding prior to operational deployment and should be revised based on the feedback obtained.

5.7 User needs and interests should inform the development of interactive or dynamic results reporting tools, including incorporation of universal design principles to ensure the broad accessibility of reporting tools and information.

Comments: See Chapter 1 for further information on universal design.

5.8 The design of online interactive tools for reporting should be commensurate with the needs and interests of the intended users in terms of both functionality and the user interface.

Comments: When integrating tools with statistical analysis functionalities, the tools should be developed to address specific reporting contexts/needs and supporting documentation and resources should likewise provide guidance about the use and interpretation of such tools in clear language. Organizations should reflect on the format of the output of digital reporting tools broadly, such as considering different frames for formatting the tools (as questions, as drop-down menu selections, etc.), and should incorporate multiple formats for results presentation (tables and graphs). All design choices should reflect best practices for accessibility for online interactive tools and should provide the same information in accessible formats (e.g., braille, other languages).

5.9 The user interface for any interactive results reporting tools should undergo substantial user testing to ensure proper functionality among all specified or known groups of intended users.

Comments: Resources for understanding the user interface and any results generated through online tool-based queries should be provided and readily accessible to users.

5.10 Procedures should be established to ensure informational and interpretive materials to support results reporting are available to anyone who accesses or generates digitally based reports.

Comments: Such material should be written in a manner clearly understandable to consumers of the test results. User input regarding resources for support interpretation and use should be gathered. Interpretive resources should be made available in the appropriate digital format (including accessible formats) when they are ready, and the resources should be maintained and kept current to the greatest extent possible.

Guidelines for Quality Control in Score Reporting

5.11 Structured quality control (QC) procedures should be prepared in advance and documented and implemented to ensure the accuracy of reported scores.

Comments: A checklist of all the QC procedures should be prepared. For each procedure, a detailed explanation of the activities and rules (when to alert and what to do) should be defined.

5.12 An automatic system should be constructed to increase scoring efficiency and accuracy, and this system should be reviewed by experts trained to identify irregular data.

Comments: Examples of irregular data include Divergent (Lower/Higher) scores in specific test forms, divergent scores in specific test locations, divergent gaps between subscores in the individual or group level, and divergent gains (from previous exams) for specific examinees.

5.13 If early reporting is required while some QC processes are still pending, it should be made clear the preliminary scores are tentative.

Comments: If irregularities are identified, an in-depth and careful examination of the scores should be carried out.

5.14 If scores are capable of being changed, measures should be taken to prevent tampering or unauthorized adjustments of scores, including an audit trail or logging system that records original scores and any changes. The audit trail should be protected from all changes and only available to authorized users.

Comments: See also Chapter 6. Data Management.

Guidelines for Maintaining Confidentiality of Score Reporting

5.15 Policies and procedures should be developed relating to the confidentiality of scores.

Comments: These policies and procedures should define who can have digital access to scores within and outside the assessing organization. Access should be restricted on a need-to-know basis, so only a small number of people have access to scores, including situations where someone other than the test taker has control over the score reporting (e.g., an employer) or where the test taker has given permission to share the score report with others (e.g., a certification body, a potential employer). These policies should also define which score information is shared with specific stakeholders.

5.16 Anyone who has digital access to non-public scores and any information that can identify test takers should be bound by a confidentiality agreement.

5.17 Where digital reporting efforts make use of secure, login-based portals, testing organizations should have procedures in place to ensure data available to specific login credentials are appropriate for the role/level of the user.

Comments: See also Data Privacy guidelines (Chapter 9) and Quality Control guidelines (Chapter 6).

5.18 When group-level interactive reporting tools are available, data privacy mechanisms such as minimum display thresholds or statistical sampling approaches should be implemented according to explicit access levels assigned to specific data user roles.

Comments: These privacy mechanisms will help prevent individual test taker results from being identified through progressive narrowing of the sample with drop-down menu selection or other data selection techniques for users who are not authorized to have such access. See also Data Privacy guidelines (Chapter 9) and Data Management guidelines (Chapter 6).

5.19 Scores should be communicated to assessment participants, or other entities authorized to receive test scores, in a way that ensures only the intended recipient receives them and the transmission is secure.

Comments: If scores can be accessed in a digital platform, the platform should have secure authentication to identify the user, either single sign-on from another system or a strong password using industry-standard mechanisms. Where possible, assessment organizations are encouraged to consider the use of multifactor authentication. If the assessment participant is sent scores by email, encryption of such scores, either by using transport email transmission or by sending them in an encrypted file with the password communicated layer security in in a way other than by email, should be used. When included in a certificate or other formal communication of results, measures should be put in place (e.g., cryptographic verification) to prevent tampering with certificates. Organizations that use scores for specific purposes (e.g., admissions, selection, etc.) should receive the scores directly from the assessing organization and not from test takers. Intended assessment participants may include parents and guardians in cases where minors are assessed.

5.20 Systems that hold test scores should have information security principles in place that are congruent with ISO 27001:2013, or similar security standards (e.g., the US NIST Cybersecurity Framework). Where possible, certification against ISO 27001:2013 is desirable.

Comments: Other approaches, such as self-certification, may be applicable in some situations.

5.21 When databases or data files of anonymized data are made available to users for import into external statistical software for analysis, details about the data, including what variables are included and excluded, should be provided.

Comments: When data files for external analysis are made available for public use, all data protection strategies relating to anonymizing data and ensuring privacy should be implemented. See also Data Privacy guidelines (Chapter 9) and Quality Control guidelines (Chapter 6).

6. DATA MANAGEMENT

Background

Technology-based assessments ("TBAs") generate important data that must be managed and maintained securely and accurately to assure the integrity of scoring, reporting, and other dependent processes. This chapter discusses issues and outlines guidelines for assessment data storage, maintenance, security, and the integration of assessment data with other systems. Other chapters in these *Guidelines* on test security (Chapter 8) and data privacy (Chapter 9) are especially pertinent to this chapter. The guidelines in this section reference various technology standards that are likely to evolve with rapidly changing technologies and, thus, should be checked for the latest versions.

Data Governance

Data are a critical asset for organizations that develop and deliver TBAs and are fundamental to the value of assessment. Throughout the assessment lifecycle, data play a critical role in the content, delivery, scoring, and reporting of assessments. Important considerations for data storage include: (1) data governance policies and practices that hold particular relevance for TBAs; (2) data architecture that hold particular relevance for technology-based assessment (TBA), in particular in relation to item banking; and (3) ongoing pursuit of mature data strategies within a context of rapidly changing and improving technologies for data storage, management, and analytics.

Data Maintenance, Integrity, and Security

It is important that TBA is conducted in a way that data captured during the assessment process are recorded accurately and securely. If data are not captured accurately or are lost, breached, corrupted, or tampered with, the credibility, validity, and integrity of the assessment can be compromised. It is also important to ensure assessment responses are retained in the event of a technology failure for continuity, record-keeping, and auditability.

Technology threats to the integrity of data are many (e.g., bugs or errors in the technology, mistakes when new software releases are made, connection failures). Overloads due to high usage (e.g., a large number of assessment participants starting or submitting the test at the same time) can also occur. Process failures, poor system architecture, human error, and attempts by bad actors to disrupt or share the data are also possible. Thus, it is important to manage and mitigate these threats and the risks they pose. Cloud-based technologies, platforms, and services are often used to address data integrity, scalability, availability, and reliability challenges.

Integrating Assessment Data with Other Systems

Assessment often occurs within an ecosystem of learning, achievement, and analytics. Results of TBA are often used in combination with data from other systems within the ecosystem for a variety of purposes. Some of these purposes include recommending or prescribing curriculum and learning/instructional content in adaptive instructional systems, grouping learners for instructional interventions, aggregating results across individuals for accountability or program evaluation, integrating data across products to inform an overall student profile for targeted intervention, adding to a graduate school admissions portfolio, and awarding digital credentials/badges, licenses, and certificates. These applications require a robust data infrastructure in which assessment data can be integrated with data from other systems to enable accurate inferences from learner interactions and to inform future learner interactions through the real-time, personalized delivery of instructional, practice, or assessment content.

The interpretation of an assessment result should be consistent with the specific purpose and use for which an assessment (or learning game or personalized lesson) was designed. Assessment interpretations are at risk of being distorted in systems without proper guidance on how the assessment data and results may be used and under what conditions. In addition, to achieve the purposes described above, algorithms are applied across data from different systems for predictive analytics. When algorithms use assessment data automatically, it is critical to maintain validity by ensuring assessment data carry with them information that allows receiving systems to use and interpret the assessment data in a manner consistent with the design and validity evidence. Further information on integrating assessment with learning may be found in Chapter 1, and information on the use of AI algorithms for data analytics is found in Part IV.

Management and governance of data to enable such use cases is challenging because it may bridge multiple entities, requiring explicit technical protocols for communications and data sharing among disparate systems. Interoperability standards, including of data formats, help to formalize and standardize these handshakes, reducing the technical work required to integrate multiple systems. The IEEE Learning Technology Standards Committee, IMS Global Consortium, the Common Education Data Standards, and the Ed-Fi Alliance are among the organizations that invite members to collaborate on standards and reference architectures. More information on interoperability may be found in Chapter 3.

As not all ecosystems will be operated under a single overarching data strategy and governance, it is therefore incumbent on assessment organizations to design data exchanges for appropriate interpretation by downstream and integrated systems.

Guidelines for Data Management

Guidelines for Data Storage

6.1 Data architecture, modeling, and solution design should be conducted in collaboration with users of the item bank or other assessment data.

Comments: These users include but are not limited to content developers, assessment designers, psychometricians, research and data scientists, and technologists designing upstream and downstream systems.

6.2 Data models should be designed to address the management of different versions and stages of assessment content and allow for extensible metadata describing attributes of item, media, and shared assessment stimulus.

Comments: Recommendations for data management and models include the following:

- *Metadata may be populated either by content developers or by automated systems (e.g., AI classification of items) and should include statistical, context, and psychometric characteristics, as well as references to support the correct answer.*
- *Use a controlled, common vocabulary (explicitly allowed terms) to facilitate indexing, categorizing, tagging, sorting, and retrieving of data when possible.*
- *Capture data required to establish diversity and inclusion requirements of the assessment (such as gender and ethnicity content tags).*
- *Capture data required to meet accessibility requirements of the assessment (e.g., Alt text, text-to-speech pronunciation, braille files, American Sign Language video).*
- *Allow for representation of relationships among items, such as items' relationships to each other (e.g., item enemies), to shared stimulus, to parent task models, and to standards or frameworks.*
- *Avoid data redundancy in design of relationships across shared assets.*
- *Cloud technologies may be leveraged to limit the movement of data and to scale computing automatically as needed for analytics.*

6.3 Data solutions should be designed to meet non-functional requirements fit to intended use, such as query and retrieval speed, searchability, data access and privacy, and business process use of the data.

Comments: Recommendations for data model solutions include these steps:

- *Ensure data streams are suitable for recording process data.*
- *Include "data lakes" (multi-source data systems) suitable for storage of structured and unstructured data intended for analytic use.*
- *Ensure distributed data storage is suitable for high-volume data. Care should be taken to meet query speed requirements when partitioning data for distribution.*

- *Use data storage options and analysis tools that limit the movement of data for analytic purposes, when possible, for both efficiency and data privacy.*
- *Data architecture should also be able to capture all test session information, including logging system status, keystrokes, data transfers, etc. The capability to replicate the exact state of a candidate's testing experience is important.*
- *Cloud technologies can be leveraged to ensure the integrity of (physically) distributed data storage and minimize loading times across regions and availability zones.*
- *When leveraging cloud providers, ensure that they comply with the current and local data privacy legislation and rules on data localization.*
- *Sharing of data that includes test taker personal information ("PI"), between user's needs to comply with national privacy laws and regulations. Use of a written Data Sharing agreement will facilitate both legal and operational effectiveness.*

6.4 Data governance should allow for data assets to be easily discoverable and available, including documentation of data elements and data dictionaries. Governance should include access controls designed to assure data privacy and security.

Comments: Recommendations for data governance include the following:

- *Make available a catalog of data elements and descriptions (i.e., a data dictionary) to content developers, psychometricians, research and data scientists, and technologists designing upstream and downstream systems.*
- *Update data dictionaries, schemas, and access requirements to synchronize with the software via automated updates to ensure interpretability.*
- *Subject to appropriate access control, make data available to analytic tools such as statistical software, elastic compute resources, and data visualization dashboards.*
- *Take into account data privacy and test security considerations in access control design.*

6.5 Data quality should be managed commensurate with the stakes of the assessment to ensure accuracy, completeness, and consistency of TBAs.

Comments: Monitoring the following aspects of data quality is important because they may have a direct impact on the validity, reliability, usability, accessibility, and auditability of assessment results:

- *Accuracy: Checks that data conforms to the valid values established for the data field during data modeling. While accuracy of data may be hard to assess (e.g., time stamps may not accurately reflect the time if the computer system used is in the wrong time zone), attempts should be made to identify discrepancies indicating inaccurate data.*
- *Uniqueness: Checks for data duplication that may impact statistical analyses and other downstream processes.*
- *Completeness: Checks for missing data to establish the extent to which required data fields are missing data. Downstream analytics should consider the impact of missing data.*

- *Consistency: Periodic checks that data stored in multiple places within the organization agree.*
- *Lineage: Using systems to trace any data element to its source and representing any transformations made to data as it travels through the system.*
- *Timeliness: Using systems to ensure data is available when it is required for downstream use cases. This is especially important for scoring and reporting of assessment results that have deadlines that impact examinees (e.g., high-stakes tests, such as college admissions testing).*

6.6 As technologies for data storage, management, and analytics rapidly change and improve, new data-related tools and techniques should be evaluated to improve the quality, security, and timeliness of assessment data and assessment-related insights.

Comments: Several data maturity frameworks are available for organizations to assess the maturity of their data practices and policies among multiple dimensions [e.g., *CMMI Data Management Maturity or ISO 8000 (ISO, 2020)]. While these standards do not specifically address assessment data practices, they are useful to support digital transformation efforts and TBAs. As data privacy laws evolve, technologists are advised to collaborate with legal teams on an ongoing basis to ensure data storage solutions keep pace with privacy concerns.*

Guidelines for Data Maintenance, Integrity, and Security

6.7 *Data Maintenance.* Processes and procedures should be established to ensure proper maintenance of all data processed, including data backup, retention, and removal.

(a) Backup procedures should be established to ensure data are preserved at all times.

Comments: Note these recommendations for backups:

- *Backup at regular intervals, at least daily, so these are available in the event of failures.*
- *Trigger alerts to operational staff in the event automated backups fail.*
- *Test the procedure to restore backups at regular intervals.*
- *Store backups in a different location (or cloud region) to where the data are stored, so a fire or other local hazard does not also destroy the backup data.*
- *Where possible, encrypt backups to protect against unauthorized access.*
- *Where available, take advantage of cloud backup systems that can reduce the effort required for reliable backup and assure data integrity and availability. Ensure that any cloud-based hosting system/databases meet applicable privacy laws and regulations, including for compliance with security requirements.*

(b) A written data retention policy should be established, implemented, and maintained consistent with jurisdiction laws, regulations, and policies.

Comments: Data may need to be kept for a certain minimal period, such as a school year, for defensibility purposes. Data may need to be removed after a certain period, for example, in case *personally identifiable data are included in the dataset, or upon a request pursuant to national privacy laws and regulations. However, there may be legal bases for the testing organization to retain data (e.g., score challenges, lawsuits). The retention policy should distinguish between types of data that may have different requirements. Data may need to be filtered after a certain period, including, for example, removing personally identifiable information and preserving other data for analytics.*

(c) Data removal processes should be established in consideration of and compliance with applicable regulations.

Comments: It may be important to remove data reported in a technical manner to make them permanently irretrievable. Data may need to be removed from automated backups if the backup file was originated before receiving the removal request. It is recommended to thoroughly test removal processes and ensure that operational staff receive alerts if the process fails.

6.8 *Data Integrity.* Processes and procedures should be established to ensure the persistence, accuracy, and reliability of assessment responses, scores, and other artifacts and evidence of the assessment-taking process.

Comments: Methods for ensuring integrity include, but are not limited to, safe storage, audit trails, quality assurance (QA), anti-malware, capacity planning and testing, change control, and business continuity.

(a) Test taker responses should be stored soon after being made (within seconds if possible) to prevent data loss in the event of computer or connection failures.

Comments: For example, if a test taker is taking a 50-question test, and there is a failure after they have answered 10 questions, then the answers submitted should be recorded to allow analysis and/or resumption of the assessment. Some test administration systems may store encrypted responses locally and transfer the responses to the cloud at varying intervals. See 6.9 for further considerations regarding data encryption to prevent unauthorized access.

(b) A comprehensive time-stamped audit trail or log should be made of all activity conducted by the test taker and other actors in the testing process, including all changes to data stored due to such activity.

Comments: Examples of these data include test taker responses, scores, proctoring/invigilation, grading, and adjustment to responses and scores. It is important that the audit trail be stored and protected from tampering and unauthorized access and that it records errors and faults. Synchronizing clocks of all systems contributing to the audit trail is also important.

(c) All software and related technology should undergo thorough QA before being used for assessments, including data capture and scoring.

Comments: Assessing organizations are advised to seek to use good industry practice on planning and executing QA, including appropriate use of automated and manual testing.

(d) Assessment systems should include appropriate anti-malware technology to protect against malware impacting the integrity of assessment data.

Comments: Additional information may be found in Guideline 6.9 (below) and in Chapter 8.

(e) Technology systems should be tested under load to identify a maximum permissible load, and measures should be put in place to ensure that the maximum load is not exceeded.

Comments: It is important to design software and technology systems such that if systems fail due to overload, they do so gracefully without impacting the integrity of data. It is recommended where distributed systems are used that regular stress tests be conducted to verify central capacity and request participating organizations to run local diagnostics to identify potential local (bandwidth) capacity limitations.

(f) Change control procedures should be put in place for software updates or new releases to minimize the risk of integrity failures due to software updates and revert to stable versions of software when needed.

(g) A business continuity plan should be developed and regularly tested to ensure the continuity of assessment data and services.

6.9 *Technical Security.* Processes and procedures should be established to ensure technical security throughout the complete process of managing and delivering the assessment, including protection against threats to confidentiality, integrity, and availability.

Comments: Recommendations for technical security include the following:

- *Ensure security by design of product and software development processes.*
- *Use regular (automated) testing, e.g., based on Open Web Application Security Project guidelines.*
- *Separate production and staging/QA/development environments where applicable, such that developers do not have access to production data by default.*
- *Safeguard access to data on production environments (e.g., through use of a Bastion-server, only allowing access through explicit consent and for a limited time).*
- *Apply patches regularly on underlying infrastructure, operating systems, frameworks, and used components.*

(a) Encryption. When transferring data between geographically separate computer systems (e.g., assessment device and server), use encrypted channels to prevent interception and tampering. Encryption should be strong and designed to meet applicable standards and used for storing all personal data as well as confidential information.

Comments: When transferring data between computer systems (e.g., between two servers or within the cloud), and for storing data, it is helpful to use encryption. Current, applicable standards may include Federal Information Processing Standards (FIPS), ISO 18033-3:2010 https://www.iso.org/standard/54531.html. Note: These standards are likely to evolve with changing technologies.

(b) An information security incident plan should be established, including a clear process and an identified incident response team, to act quickly when incidents occur.

(c) Data security policies should be established and communicated to all relevant employees and contractors.

Comments: See Chapter 8. Test Security, for additional discussion of security polices and communication.

(d) When using non-cloud computer systems, restrict physical access to assessment materials and hardware/servers. Perform disposal of materials and hardware in a secure manner.

Comments: For example, use of keycards for physical access, physical destruction, zero-filling of hard drives, or use of a specialized third-party company. When using a cloud system, physical access and disposal are usually managed securely by the cloud provider. Verify cloud provider security standards and compliance certifications such as ISO 27001, CSA STAR, and SOC 2 attestation (or comparable).

(e) A third-party company should conduct penetration testing regularly to ensure the security measures put in place are sufficient.

(f) Policies and procedures should be established, implemented, and maintained for granting and removing access to assessment data.

Comments: It is important that access to data be granted on a need-to-know basis for persons who can be identified.

(g) Third-party review and certification of security processes and procedures should be conducted for assessment systems and data.

Comments: For example, ISO27001 certification or SOC 2 attestation (or comparable).

Guidelines for Integrating Assessment Data with Learning Systems

6.10 Assessment systems should ensure generated data travels with, or may be linked to, contextual metadata that allows a receiving system to understand how assessment data are properly interpreted.

Comments: Following are some recommended types of metadata to be generated:

- The purpose for which the assessment was designed.
- Linkage with relevant standards or competency frameworks.
- Assessment administration conditions relevant to the interpretation of assessment results.
- Hierarchical relationships present in data structures (e.g., students nested in classrooms nested in schools) so that analyses can appropriately account for them.

6.11 Receiving systems should evaluate assessment data fitness for purpose before integrating it into analytics.

Comments: It is recommended that assessment reporting systems indicate limitations in the data received, such as missing data or misalignment of assessment data with reporting classifications.

6.12 Interoperability standards designed for the transmission of assessment data and supporting contextual metadata should be implemented when feasible.

Comments: See also Chapter 3. Interoperability. When interoperability standards are insufficient for ensuring proper interpretation of the assessment data for subsequent analytics, extensions are applied, and communications to standards bodies for consideration of expanding the standard to apply and follow.

6.13 Collection and management of user data should occur in accordance with relevant laws and professional standards.

Comments: See also Chapter 8. Security, and Chapter 9. Privacy. It is important to consider test taker privacy rights and applicable privacy rules before copying personal data from one system to another. Management of user data includes linked systems capable of removing examinee data at the examinee's request and capabilities to support requests for data access and interoperability.

6.14 Assessment data should be tagged and organized in a way that allows for integration with data in other systems to support data aggregation across systems for analysis and reporting in a manner that does not violate data privacy requirements.

Comments: Methods such as anonymization and pseudonymization are used to address privacy regulations when possible. Examples of data aggregation include role (e.g., learner, educator) and/or level (e.g., activity, session).

6.15 Assessment data should be transformed and stored in a format easily consumed by analytics platforms for data analysis and reporting.

Comments: It is important that data lineage be represented and retained (Sweet, 2016).

7. PSYCHOMETRIC AND TECHNICAL QUALITY

Background

As assessment technologies advance and evolve, the principles of sound measurement remain as core concerns. Ensuring measurement quality in the era of digital assessment is at the forefront of concerns, with the aim of assuring assessment results are not adversely affected or distorted by using technology in design, delivery, and scoring. Score comparability is a specific concern when multiple testing modalities are used. Ultimately, attention turns to validation strategies and considerations for technology-based assessment ("TBA").

Quality TBAs provide an appropriate medium for measurement of the target construct without introducing construct-irrelevant variance ("CIV") in scores, construct underrepresentation, or increased measurement error. It is important that these conditions hold for all test takers for fair and equitable assessment and are supported by system documentation and evidence based on empirical research. Evidence of measurement quality may be demonstrated through empirical studies examining threats to score comparability, measurement invariance, dimensionality, and score reliability (Kane, 1982, 2011, & 2013), along with evidence of validity. The AERA et al. (2014) *Standards* outline five sources of validity evidence: test content, response processes, internal structure, relations to other variables, and consequences of testing. Assessment system documentation and quality assurance are also key for assuring the standardization of assessments and their appropriate use.

The use of technology in scoring, especially software -automated or AI algorithmic scoring and decision-making, is also a key concern. Chapter 4 provides a discussion of scoring-related issues and guidelines central to measurement quality. Further information on AI regulations is found in Part IV.

Score Precision, Comparability, and Equating

The precision of TBAs is reliant upon stable technology systems and software that is free from extraneous influences upon test performance (e.g., system lag times, cumbersome user interfaces). Thus, it is important to assess the reliability and precision of TBA scores to enable appropriate use and interpretation of scores, as well as to detect potential issues that may affect scores.

Comparability of scores resulting from assessments that use technology is a core consideration for measurement quality in many applications (Camara & Davis, in press). In general terms, the concept of comparability refers to the degree to which two or more different tests, or two or more forms of the same test, administered concurrently or at different times, or through different modes (e.g., pencil-and-paper versus computer versus tablet), can produce comparable scores (Berman, Haertel, & Pellegrino, 2020; Newton, 2010). Score comparability is not necessarily the same as interchangeable or equivalent scores and may be demonstrated in different ways that vary in level of rigor and precision. For example, comparable scores may be produced through a variety of different methods of score linking, which include concordance, prediction, and equating. Score interchangeability generally is reserved for equated scale scores, and comparability is increasingly thought of as a slightly less fine-grained comparison such as score pass/fail decisions or performance-level classifications (Berman et al., 2020; Winter, 2010). In all instances, scores from two tests are transformed to allow comparisons or predictions across the measures (Dorans, Moses & Eignor, 2010).

Traditional models of equating and item calibration require large samples of data on individual items Kolen & Brennan, 2014). Several technology-based methods have been recently introduced with the goal of reducing data collection needs for calibration processes. First, machine learning and natural language AI processing have been used with the aim of alleviating the burden of pre-testing and equating, as well as to establish score comparability for computer-adaptive testing programs. These methods seek to create items and directly estimate their difficulties without pilot testing (Settles, LaFlair, & Hagiwara, 2020). Second, assessment design and automated item generation methods have reconceptualized alternate form development, with items as tightly controlled instantiated units within larger task- and item-model families with template-driven item development of automated computer software that employ item-cloning templates or shells (Gierl & Lai, 2013; Luecht & Burke, in press).

Some innovative TBAs, such as game-based assessments and complex performance assessments, may not lend themselves to the same level of comparability associated with more traditional assessments. In such instances, it is important that tradeoffs between score comparability and other objectives of the assessment design be balanced and documented (Mislevy, Corrigan, Oranjc, DiCerbo, Bauer, von Davier, & John, 2016).

Assessment delivery modality and technology differences may be of particular concern, especially in high-stakes testing. For example, switching from paper-and-pencil to CBT to online Internet-based testing using different devices and systems could introduce unintended differences in score interpretations. The transition to a new modality or concurrent use of different assessment modalities warrants examination via empirical research (Sireci, 2005; Winter, 2010; Lottridge, Nicewander, Schulz, & Mitzel, 2010). In cases where multiple modes of administration are used (PC, tablet, smartphone, or paper-and-pencil), mode effects and other factors (e.g., environmental effects, test security, technology disruptions, test taker experience) should be empirically studied to address the potential impact on score interpretations.

Measuring Change and Growth

Technology enables the linking of data from different assessment and learning systems and enhances the possibilities to measure change and growth in test taker performance (see Chapters 3 and 6). A major benefit of TBAs is that they are not only able to tailor a test to meet the instructional goals for a test taker, but the testing schedule may also be coordinated to further the goals of the assessment. The increased flexibility to design instruction and evaluate its outcome potentially increases the validity of measures of achievement and growth in service of instruction, especially as evaluation of achievement is made in a timely manner in support of learning. Moreover, technology-based platforms for sustaining learning and assessment may extend to many contexts and frameworks proposed for organized learning models (e.g., Almond et al., 2012), as well as systems seeking to integrate learning, assessment, and evaluation more fully (e.g., Gordon, 2000; von Davier et al., 2017).

The overall information gathering potential of TBA platforms promises to increase the capacity to extensively track test taker performance outcomes and testing conditions (all with test taker consent; see Chapter 9). Efficiency in applying advanced learning assessment models is also increased. For example, using AI to adaptively drive the time-sampling of multiple input and output domains enables continual capture of the examinee's instructional and learning progress. AI-driven monitoring subsystems may then return reflexively as feedback to further steer measurement, testing, and decision-making. Additional information on AI-driven systems is found in Part IV. Integrated assessment and learning databases also enable post hoc modeling to develop inferences about growth and change and potentially overcome information gaps that help reduce threats to validity (Campbell & Stanley, 1966; Cook & Campbell, 1979).

Validation of Technology-Based Assessments

Validation of assessments refers to compilation and evaluation of evidence regarding the use of test scores for their intended purposes. As mentioned in Chapter 1, the AERA et al. (2014) *Standards* state validity "refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (p. 11). These *Standards* propose five sources of validity evidence that can be used to evaluate test score interpretations and uses. These five sources are validity evidence based on (a) test content, (b) response processes, (c) internal structure, (d) relations of test scores to other variables, and (e) consequences of testing. All five sources are helpful for comprehensive validation of TBAs.

In establishing validity evidence for TBAs, CIV and construct underrepresentation are key considerations, as noted in Chapter 1. The term "construct" is used to describe the knowledge, skills, abilities, or other personal attributes measured by an assessment. Technology is often used to increase construct representation by allowing measurement of knowledge, skills, abilities, and attributes that were impossible or very difficult to measure without technological innovation. At the same time, there is a concern that TBAs may measure irrelevant test taker characteristics such as computer literacy that lead to inaccurate measures of the constructs targeted by an assessment. Thus, issues of maximizing construct representation and minimizing CIV are key focus areas in the validation of TBAs.

The five sources of validity evidence are helpful for evaluating construct representation and the potential presence of CIV. Descriptions of specific actions that can be done appear throughout these *Guidelines*, and readers are referred to the AERA et al. (2014) *Standards*, as well as seminal sources describing validity evidence based on test content (e.g., Martone & Sireci, 2009; Sireci & Faulkner-Bond, 2014; Webb, 2007), response processes (e.g., Padilla & Benitez, 2014; Zumbo & Hubley, 2017), internal structure (Wells, 2021); relations to other variables (Kane 2006; Schmidt, 1988), and testing consequences (e.g., Lane 2014).

As mentioned earlier, one validity issue particularly relevant to TBAs is *comparability*. In its most general sense, "comparability" refers to the degree to which test takers' scores on a test can be meaningfully compared. This issue is relevant when assessments are delivered on different devices (e.g., laptops, desktops, tablets, handheld devices), different digital platforms (browsers, operating systems), across different languages, or different forms of a test. In the event that CIV in test taker scores is introduced by any of the foregoing factors, comparability will be affected. Chapter 11 of these *Guidelines* addresses considerations for assessment in different languages (see *Translation and Adaptation*).

With respect to validation, the AERA et al. (2014) *Standards* state, "A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses." (p. 21). Thus, validation of TBAs ideally involves a compelling synthesis of various sources of validity evidence to support the intended uses of test scores. With the goal of a comprehensive body of evidence to support the use of a test for the entire spectrum of examinees tested, we offer the following guidelines.

Guidelines for Psychometric and Technical Quality

Guidelines for Score Precision, Comparability, and Equating

7.1 TBA delivery and standardization practices should be defined and documented in sufficient detail to support efforts to mitigate threats to measurement quality.

Comments: *Threats to measurement quality include CIV in scores, construct underrepresentation, or increased measurement error due to the use of technology. Hardware and software infrastructure (e.g., system architecture) is part of the standardization of testing conditions. Chapters 4 and 6 provide details regarding data and scoring quality measures, and Chapter 11 provides information regarding test taker preparation to help mitigate CIV.*

7.2 Evidence of measurement precision (reliability) should be provided throughout the range of the scale to support its uses and interpretations.

Comments: Many TBAs use adaptive technology where the concept of a reliability estimate for a set of items or a test form does not apply. Test information functions and conditional standard error curves are appropriate for reporting measurement precision for these types of TBAs. Where feasible, a testing organization should identify and account for major sources of measurement error and provide evidence of reliability and measurement precision for relevant subgroups of examinees. Traditional reliability estimates may be reported for linear test forms, regardless of test administration mode.

7.3 When TBAs involve multiple test forms, appropriate equating methods should be used to ensure the equated forms measure the same construct at a comparable level of difficulty and precision.

Comments: Further recommendations for equating under various conditions and models may be found in Dorans and Puhan (2017) and Kolen and Brennan (2014). Approaches to linking and equating scores for TBAs should be appropriate for the specific application, intended claims, and use case, e.g., when a test is administered using multiple modalities, devices, or administrative conditions. Additional use cases may include when two or more forms of the same test are produced; when some form of computer-adaptive testing is implemented; when design differences exist that could impact constructs or performance (different timing, different response options, accommodations, etc.); and when a test is redesigned or updated (changes in blueprint, item types, construct, timing).

7.4 Claims of assessment score comparability should support score interpretation across different technologies, devices, and administrative conditions, as well as different test forms and items, where relevant.

Comments: When the purpose of a TBA requires score comparability across test variations, score equating, or another form of linking may be needed. Evidence the construct is measured comparably for relevant groups of test takers should be provided (e.g., measurement invariance studies, differential item functioning analyses, test event-level alignment analyses); see also Chapter 10. Test content, timing, rendering, responding, and cognitive processes could be influenced by the technology, devices, mode, platform, and other administrative and environmental conditions. The degree to which these conditions affect score comparability should be studied. Evidence supporting score comparability may entail collecting multiple sources of validity evidence. The test developer has a responsibility to provide evidence to support any claims of comparability, while the test user is responsible for ensuring additional variations are not introduced during or subsequent to the test administration.

7.5 Intended variations of assessments should be defined, and documentation should be provided regarding how measurement quality is maintained or enhanced.

Comments: *In some cases, comparability is not intended and the degree to which variations affect test score interpretation should be clearly specified.*

7.6 Psychometric evidence for TBA score comparability should include an evaluation of the properties and differences in the shape of the score distribution, reliability, and standard error of measurement.

Comments: The effect of administration mode may be examined at both test and item levels. Claims of score equivalence could consider distributional equivalence of scores, construct equivalence, predictive equivalence, and population invariance across modes and devices. In construct equivalence, the construct across modes/devices remains the same; in predictive (correlational) *equivalence, relationships with external variables are similar (Bugbee, 1996); population invariance of linking functions across major subgroups may be examined if sufficient samples are available. When using item pools, for scores to be interchangeable from one alternate item pool to another, the item pools should be built to support the generation of forms that meet the same content and statistical specifications. Resources for evaluating comparability include Berman, Haertel, and Pellegrino (2020), Dorans (2004), Sireci, Rios, and Powers (2016), and Wang and Kolen, 2001.*

7.7 Documentation of evidence relating to score comparability or equivalence should address data sources and samples, methods, and analyses.

Comments: Recommendations for documentation include data collection procedures, descriptions of samples, methods, and analyses conducted, as well as any limitations or cautions in interpreting the results.

Guideline for Measuring Change and Growth

7.8 **Metrics and indices for measuring change or growth should be reliable and valid for their intended purposes, supported by appropriate documentation**

Comments: When inferences about test takers' changes in performance on the construct or growth are made, it is important for those inferences to be supported by evidence of reliability and validity, with reference to an underlying model of learning progress. Where no such evidence is available or where growth or change scores are considered unreliable, claims and indices of growth or change should not be provided. Interpretations of indices derived from TBAs should be supported by validity evidence. Data from assessment systems and any other integrated systems used to support inferences about growth or change should reflect, or be converted to, a meaningful scale and level of aggregation to support such inferences.

Guidelines for Validation of Technology-Based Assessment

7.9 The intended uses and purposes of TBAs should be clearly defined.

Comments: The intended uses and purposes of test scores dictate the types of validity evidence to be gathered, analyzed, and reported to justify the use of a test. Therefore, these uses and purposes should be defined for all test users, test takers, and other stakeholders, so they are clearly understood.

7.10 The construct(s) measured by TBAs should be clearly defined.

Comments: Test takers and consumers of test scores (e.g., teachers, employers, researchers, certification bodies.) should understand what a TBA measures. Clear definition of the construct measured should include descriptions of the content and cognitive domains measured on educational tests; the knowledge and skill domains measured by credentialing exams, the *personality dimensions measured on personality assessments, attitudes measured on surveys, and so forth. Test specifications that describe these areas and domains and serve as operational definitions of the constructs measured should be made available to test takers and those who interpret test scores. Confirmation a TBA is measuring its targeted construct(s) is a fundamental step in validating the assessment.*

7.11 Validity evidence should be provided to support the intended uses of TBA scores.

Comments: Validation of TBAs should begin with considerations of the types of evidence that would confirm the test is: (a) accurately measuring the intended constructs; and (b) not measuring unintended constructs. A single study is not likely to provide sufficient evidence to support the use of a test for its intended purposes. Rather, multiple sources of validity evidence should be synthesized into a coherent validity argument that supports test use. The types of supporting validity evidence and strategies may vary with different assessment purposes and contexts; e.g., education, employment, clinical, and credentialing assessments may focus on different constructs, predicted outcomes, or alignment with content domains.

7.12 Validation of TBAs should confirm the infrastructure required to deliver and interact with the exam does not impede test takers' performance.

Comments: A comprehensive validity argument for TBAs should confirm test takers understand how to interact with the system to successfully access the test and provide their responses. Computer literacy should be ruled out as a source of CIV. In addition, the user interface should be evaluated to ensure it is not causing undue stress or cognitive load for test takers to successfully receive and respond to test items.

7.13 Validation of TBAs should consider the diversity of the test taker population and the degree to which interpretations of test scores are consistently fair across groups.

Comments: Test takers are likely to differ from one another in many ways, such as gender, race/culture, socioeconomic status, disability, age, and other personal characteristics. Studies of invariance at the item level (i.e., differential item functioning) and test level (e.g., differential test functioning), as well as criterion-related validity of test scores (e.g., differential predictive validity), can help evaluate potential aspects of bias and unfairness across groups of test takers. Qualitative analyses such as think-aloud protocols or interviews may also be illuminating with respect to fairness and how testing programs can be improved to be maximally inclusive. Consideration of diversity and fairness begins at the earliest stages of test development. Culturally sustaining test development practices can improve the validity across all test takers by ensuring test content and contexts embrace the totality of cultural variation within the tested population (Randall, 2021).

7.14 Validation of TBAs should ensure the time limits established for the test are clear and reasonable.

Comments: Test takers should have sufficient time to complete all test items and demonstrate their full potential with respect to the constructs measured. If speed of response is explicitly part of the construct measured, the degree to which the test measures this construct should be clear in the construct definition and clearly communicated to test takers. Test takers should be instructed on how to best use their time in taking the test and how speed of response will affect their scores. The degree to which test takers understand timing and scoring rules can be important validity evidence.

7.15 Validation of TBAs should be conducted on a periodic basis to (a) confirm use of the test continues to be justified by evidence and (b) to improve the testing program.

Comments: Validation of TBAs should be both formative and summative and should be conducted on a regular basis to acknowledge the changing nature of the assessment and the examinee population. It is likely validity studies will point out strengths and areas for improvement in a testing program. Evidence pointing to areas of improvement provides formative information that can improve validity. Nevertheless, a summative conclusion ultimately needs to be made that the test is justifiable for its intended purposes. That conclusion should be updated based on new validity evidence as a testing program progresses.

8. TEST SECURITY

Background

Security is a long-standing concern in high-stakes testing and represents an especially important factor for technology-based assessments ("TBAs") because they may be deployed in a wide range of modalities and settings. Test security is important because the validity of test scores relies on the requirement that each test is taken in accordance with proper procedures, following security guidelines and rules, and where the test taker's identity has been verified. Any security failure can impact the validity of test scores and, therefore, the integrity of the testing program. Test security applies to both test content and results, as well as to test takers' personal information ("PI") collected and used by the testing organization (discussed in Chapter 9. Privacy; see also Chapter 6. Data Management, for guidance on information security, including use of NDAs/Confidentiality Agreements).

The guidelines in this chapter are intended to assist testing organizations and test users in safeguarding against potential security threats and risks to testing programs and enable them to focus resources on the most important vulnerabilities and strategies to protect the testing organization's assets. This chapter provides basic information on test security threats, risks, and protective strategies to assist in the application of the *Guidelines*.

Security Threats and Risks

Threats do not automatically mean security breaches but may result in a breach if not dealt with capably. For example, the threat is, "My test items can be stolen using a camera," whereas a breach is, "I have evidence that someone stole my items using a camera." There is a substantial difference between these two scenarios. For a threat, no damage to the test or the program has yet happened, and if the threat is identified and addressed with a mitigation strategy, a breach might never occur. Categories of threats were developed in the ATP publication entitled, *Assessment Security Options: Considerations by Delivery Channel and Assessment Model* (ATP, 2013), and in *The ITC Guidelines on the Security of Tests, Examinations and Other Assessments* (International Test Commission, 2014).

Threats to score validity are listed in Tables 8.1 and 8.2, which are organized into two types of test fraud: cheating and theft. Cheating threats have the singular goal of increasing test scores, thereby introducing construct-irrelevant variance that would undermine validity. Theft threats, on the other hand, are driven by the goals of harvesting or pirating test content so that the content may be used, shared, or sold to others for monetary gain. Successful test security efforts depend on awareness of these threats, evaluating the respective risks to the program, and putting in place solutions to mitigate the risk.

Risk is defined as the likelihood of an event multiplied by the potential negative impact from the event. Not all test security threats involve the same level of risk. Some threats are rare but would have a significant negative impact if they occurred. Others may be continuously present but have relatively small impact. And some--those with the highest risk--are both very likely and very damaging when they occur. It is important for programs to evaluate each threat for the risk it poses, as the same threat could pose a different risk for different programs. Once such a determination is made, security resources can be applied to put in place solutions for those threats that carry the highest risk. As it directly applies to monitoring security risks, consideration should be given to evaluating all testing technology against the requirements of ISO 27001, *Information Technology – Security techniques,* or other credentials to verify conformance with strong security practices (e.g., SOC II audit). See ATP's Privacy in Practice Bulletins (2019-2020). See also the *Standards for Educational and Psychological Testing* (AERA et al., 2014, Chapter 6).

Table 8.1. Categories of Score Validity Threats Due to Cheating

Using pre-knowledge about the test
Receiving expert help while taking the test
Using unauthorized test aids or assistance
Using a proxy test taker
Tampering with testing software or stored test results
Copying answers from another test taker during the test
Manipulating testing rules

Table 8.2. Categories of Score Validity Threats Due to Test Content Theft

Stealing test files before, during, or after an exam
Stealing questions using digital photography
Stealing questions by capturing test content electronically
Memorizing test content for subsequent recording or sharing
Transcribing questions verbally into a recording device
Obtaining test material from a trusted insider
Manipulating testing rules

The outcome of the risk analysis will be different for each testing program and context. For example, the biggest risks for U.S.-based K-12 testing are likely to differ from those of occupational certification programs. These differences mean the variety of protective solutions will end up being different for each program (Wollack & Fremer, 2013). It is possible, even likely, that not all guidelines described in this chapter will be appropriate for a particular testing organization's security needs.

Test Security Strategies

The goal of test security is the protection of data. From a validity perspective, the most important organizational assets are the integrity and meaning of test takers' scores resulting from assessments within a specific program. It is from these data that important decisions are made. For many stakeholders, protecting the confidentiality of test content is a primary method for ensuring the integrity and meaning of test takers' scores and is also important given the value of the tests to the program and the cost to replace them. Test takers' PI, as well as test takers' scores and results and other confidential information, must be protected as well. There may be other assets on which a testing organization may wish to expend resources to protect.

Three general sets of solutions are needed to protect testing assets and can be considered as equal in importance. These are prevention, deterrence, and detection/response.

Prevention. Solutions designed for prevention make it less likely that a threat is able to turn into a breach, and if it does, the solution design should limit potential damage. One example is the well-known characteristic of adaptive tests to reduce the overall exposure rates of test items. If items are exposed fewer times during test administrations, the opportunities for capture are reduced. Another example of prevention is the randomization of the order of possible answers presented for a multiple-choice item. The random ordering of response options makes cheating more difficult when attempting to copy from an adjacent test taker or using a published cheat sheet with specific response option labels. (Note: Prevention solutions are intended to reduce the likelihood that threats progress to the level of a breach).

Deterrence. Solutions designed to persuade a person that cheating or harvesting items is wrong and not worth the effort can serve as powerful deterrents. At the heart of deterrence are effective communication as to the rules, their enforcement, the certainty of detection, and the related consequences of breaking the rules. Requiring every test taker to read the rules and sign an agreement to abide by them is an example. (*Note:* Deterrence solutions are intended to reduce the likelihood that threats progress to the level of a breach).

Detection/Response. Appropriately designed joint detection/response solutions are intended to identify the occurrence of a breach (detection) and immediately implement previously designed associated actions (response). Detection without response is typically ineffective. Detection examples include use/monitoring of a tip line, web monitoring for test content, use of watermarked items, or data forensics, as well as many other methods that prompt responses to cheating and breaches. Early detection combined with a prompt response may help to contain and mitigate damage from breaches.

The guidelines set forth in this chapter are organized according to the fundamentals of test security discussed above. Not every solution will be appropriate or useful for every testing organization, and many testing organizations have developed their own written security policies and written plans for responding to security incidents/data breaches. It should be remembered as well that the guidelines are general statements, with an example or two to help clarify them. The precise nature of a specific solution, which might combine multiple guidelines, will be unique for each program based on their needs, resources, and the nature of each threat.

Guidelines for Technology-Enabled Test Security

8.1 A testing organization should develop and follow a written security plan, to be updated at least annually. The security plan should address the following areas: relevant threats and risks; roles and responsibilities for managing and administering the program, including critical security incidents and confirmed data breaches; test user, nondisclosure and other agreements; test taker rights and responsibilities; procedures for challenges/appeals; training; and communication.

Comments: Suggested elements of a security plan include the following:

- *Create a list of relevant threats, a risk analysis process, and the principles of and steps to create specific solutions.*
- *Identify and describe all test security roles and responsibilities between all parties participating in the testing services.*
- *Specify the organizational roles responsible for managing and administering the plan; these individuals will evaluate threats and the risks associated with those threats.*
- *Specify procedures for logging security incidents.*
- *Include confidentiality/nondisclosure or similar types of agreements to be signed by participants in the plan.*
- *Specify the individuals responsible for establishing, managing, and evaluating the solutions.*
- *Specify the rights and responsibilities of test takers as they relate to test security incidents, including the need for individuals to take responsibility for being aware of security issues during preparation for and administration of a test (including individuals who receive accommodations and accommodation providers, such as readers or translators).*
- *Establish an appeals process for challenging test results and communicate it to test takers.*
- *Provide security training for all individuals involved in security efforts.*
- *Adopt, implement, maintain, and disseminate a list of specific test security rules.*
- *Fund security efforts appropriately so that protective and continuous solutions can be put in place, including contingency funds, available in the event of a breach, for investigating and mitigating the effects of the breach.*
- *Include a detailed action plan in the event of the detection of a threat or a breach (see 8.8 below).*
- *Cover legal issues associated with managing test security (e.g., reference and promote adherence to applicable breach reporting laws, privacy laws, and copyright laws).*
- *Address and investigate potential and actual breaches (see 8.7).*
- *Require the review and approval by key stakeholders annually.*

8.2 A testing organization should continuously analyze the risk of cheating and theft threats and adopt, implement, and maintain appropriate solutions for those threats that carry the highest risk.

Comments: For example, in the event that data forensics results identify a marked increase in the passing rates on a test, a search of the Internet for brain dump sites may be conducted to identity potential threats. Another example might be in the case of a tip line that enables test takers to report instances of proxy testing at a certain test center, which leads to an investigation with data forensic analyses and review of videos of test taking at that location.

8.3 **A testing organization should evaluate on an ongoing basis the technology it uses in test development, test administration, and at other stages of a test's lifecycle, as well as in the storage, transfer, retention, and destruction of test data and test taker PI, to make sure it is providing the desired protection of the testing organization's assets and is free from vulnerabilities that would put test scores and other testing data and test taker PI, at risk**.

Comments: It is important for organizations to keep up to date on changes in technology to ensure they take steps to improve and enhance their security solutions.

8.4 **A testing organization should adopt, implement, and maintain measures to prevent test fraud throughout the test development and administration lifecycle. These measures should include design and development of the types, formats, and features of items and tests; training of (internal and external) participants in test development, test delivery and administration; and design of test assembly, test administration locations, and software (see Chapters 1 and 2).**

Comments: Following are some suggested measures to prevent test fraud:

- *Design the types, format, and features of* ***items*** *to prevent high-risk threats. For example, randomizing options for multiple-choice items will mitigate copying and help to make item content unpredictable, frustrating attempts at successful content theft. Development and test administration systems should be capable of building and using secure item formats.*
- *Design the types, format, and features of* ***tests*** *to prevent high-risk threats. For example, use multiple equivalent test forms or computer-adaptive testing in order to reduce overall item exposure rates. As another example, randomize items on the test to mitigate copying and answer key sharing. Make sure development and test administration systems are capable of building and using secure test formats.*
- *Train item writers, reviewers, translators, editors, and others involved with test development, on the need for confidentiality, restrict access to the content that is necessary for them to carry out their assigned tasks, and remove access once their tasks are complete.*
- *Train, restrict access, and ensure qualifications of individuals involved in providing accommodations for test takers (e.g., readers, translators); see National Center for Educational Outcomes, 2015.*
- *Use confidentiality/nondisclosure agreements with all personnel who have access to items/test forms or other sensitive information (test takers' PI and test scores) (see Chapter 1).*
- *Design test administration locations to deter security problems. For larger sites, seating arrangements should prevent collusion between test takers. Make sure that no one except*

the test taker can view test content on his or her screen (e.g., use screen protectors). Test content and results stored at the testing location should be strongly encrypted and have adequate access control measures. Technology used for test administration and/or remote proctoring should prevent access to prohibited digital resources (e.g., using a lockdown browser).

- *For remote delivery of tests, including the use of proctoring services, adopt, implement, and maintain specific procedures and protocols around the use of such delivery and proctoring to ensure test security to the maximum extent possible.*

8.5 A testing organization should adopt, implement, and maintain authentication technology and procedures to ensure only the authorized individual is sitting for the exam.

Comments: Suggested authentication steps include the following:

- *Use appropriate and secure identification documents, preferably more than one.*
- *Use reliable, private, and safe technologies (which may include biometric measures) to enable matching the test taker's identification at registration with identification used at the testing event.*
- *Adopt and implement appropriate best practices for using biometrics to collect sensitive biometric data to comply with applicable privacy/security laws and regulations.*
- *For internal testing within an organization, it is helpful to require takers to sign on to the computer with their organizational credentials using single sign-on since individuals are less likely to share those credentials with others.*

8.6 A testing organization should adopt, implement, and maintain measures to deter test fraud. These measures should include, but not be limited to: communication to test takers requirements, responsibilities, procedural rules and rights; disclosure of prevention and detection measures; use of agreements; and, when appropriate, copyrighting content.

Comments: Following are some suggested measures for deterring test fraud:

- *Provide test takers with the requirements, procedures, and reasons for honesty. Provide opportunities for test takers to agree in writing or digitally with those requirements.*
- *Communicate the testing organization's test security rules and procedures to test takers and others and explain the consequences for breaking those rules. Establish and explain the appeals process.*
- *Describe, generally (or, if comfortable, with some level of detail), the testing organization's prevention and detection measures. Make it clear that cheating will not be tolerated and that cheaters will likely be caught and penalized.*
- *Use written agreements (e.g., test taker forms, nondisclosure agreements) to make sure that test takers are aware of the seriousness of the commitments they make.*
- *Copyright items and tests when and where possible, and make sure that test takers and others are aware of copyrighting efforts and the penalties for theft or infringement.*

- *In communications to test takers and others, emphasize that an effective proctoring presence, either on-site or online, is in place to detect attempts at test fraud.*
- *When required, give notice and an adequate explanation of the use of artificial intelligence in proctoring, administration, or scoring of the test; see Chapter 9 (Data Privacy) and a discussion about AI regulation in Part IV.*

8.7 A testing organization should put in place measures to detect and report cheating or content theft and respond to them as quickly as possible. These measures may include data forensics, monitoring Internet sources for disclosed content, monitoring the test taker during the test, and methods to report test fraud when observed.

Comments: Detection measures may include data forensics, which can be used to analyze the statistical properties of test results to discover unusual patterns that may be an indication of test fraud. When initial test results are taken under non-secure conditions (e.g., no proctor at all), provide a verification test under secure conditions as an opportunity to verify whether the initial exams were taken without cheating. Take whatever action is deemed necessary by the test sponsor/testing organization and is supported by the data forensics results. With respect to monitoring Internet sources and other communications systems for disclosures of test content and inappropriate discussions of tests, it is important to take appropriate actions, such as sending cease-and-desist letters and takedown notices to site owners. Create long-term and mutually beneficial relationships with site owners and Internet service providers.

Detection measures may also include monitoring the test taker during test administration to detect attempts to engage in unauthorized conduct. Monitoring can be done using humans and/or automated processes. Immediate and appropriate action should be taken when an incident is detected. A log of security incidents and incident response actions should be kept. The quality of security detection efforts should be evaluated on a regular basis.

Test administration technology may be used to detect unauthorized patterns of responding, such as unauthorized keystrokes to access other resources (such as control or command keys, escape, or print keys). When prohibited response patterns are detected, appropriate alerts can be immediately issued. Maintaining an audit trail will provide further evidence.

Providing a telephone, email, or web page tip line for test takers and other stakeholders are effective ways to report fraud. It is helpful to make the tip line available for individuals to report to the assessment organization in the event they discover or hear about a breach or threat. Providing clear directions for when a tip is received is also helpful, such as an immediate review of the tip by the members of a security committee.

Responding when detection systems indicate that test fraud may have occurred typically includes an investigation to corroborate preliminary evidence, which may include proctor observations, video records, data forensic results, or other information.

8.8 A testing organization should develop, implement, and follow a written incident response plan ("IRP") to prepare for, prevent, detect, report, and remediate any security incident or potential data breach.

Comments: An incident response plan should include roles and responsibilities, methods and rules for detection, logging incidents, procedures and policies that address consequences of cheating or test fraud once detected, procedures and policies addressing security of test accommodations, and disciplinary actions or penalties associated with breaking each type of test security rule. See ATP Privacy in Practice Bulletin #5 (2019) and #10 (2020) for further details regarding incident response planning.

9. DATA PRIVACY

Background

Building from the previous chapter on security, data privacy has become an equally serious concern in testing and represents an especially important factor for technology-based assessments (TBAs). Privacy is an important ingredient in protecting the integrity of an individual test product or of an entire testing program; the reputation of an organization can be affected (positively or negatively) by the testing organization's ability to provide required protection of individuals' personal information (PI) and in some situations, of test data and outcomes. Privacy requires the complementary use of test security and information security so that PI collected and used by the testing program is adequately protected (see*, also,* discussion in Chapter 6. Data Management and Chapter 8. Test Security).

Legal requirements in the European Union, Brazil, Canada, China--and some state laws in the United States--reflect an emerging broad international consensus about the need to protect PI and provide individual rights regarding PI. This chapter includes guidelines all TBA programs should follow as a matter of good practice in complying with applicable laws and regulations. Depending upon the jurisdiction(s) in which a testing organization operates, there may be varying privacy requirements imposed by law. Often then, an organization will need to develop an appropriate and proportional balancing between administering its tests, protecting its intellectual property (IP) and its other legitimate interests, and test taker privacy. Where a testing organization operates in different parts of the world that have different privacy requirements, the organization will need to address those applicable requirements appropriately.

Compliance with privacy laws and regulations involves an evaluation of what PI is collected and used, the purposes for which it is used, where and how it is stored, and with whom it is shared. It also matters whether the testing organization makes the decisions on what PI is collected and used (i.e., is the controller of the data), or the organization is merely processing test taker PI for the entity that is the controller (i.e., is the processor of PI at the direction of the controller). In some situations, a testing organization may actually serve in both roles. If an organization is only operating in a single jurisdiction, adopting a thorough approach to protecting privacy can be relatively straightforward; however, if the organization operates across multiple jurisdictions, adopting a comprehensive set of privacy practices can become very complex. Generally, a multi-jurisdictional organization will prefer to have a single uniform privacy policy to follow rather than attempt to target its compliance efforts to each specific jurisdiction on a case-by-case basis. Thus, the testing organization must balance specific legal requirements and seek to implement an approach resulting in a "reasonably defensible" privacy scheme. A testing organization should try to adopt its privacy plan early in the process of establishing its procedures, to avoid needing to add procedures that could complicate a smooth, effective process.

As part of their operations, TBAs gather PI (e.g., name, address, email address, identification information, including payment information, that identifies the individual or is associated with a specific individual), as well as information about aspects of individuals' personality, ability, or competence and share such information with multiple organizations (e.g., service providers) and locations as part of standard registration, scoring, reporting, processing, and research activities. In clinical applications of assessment, test data may be considered electronic health or medical records that are legal protected.

In the environment of TBAs, issues related to the collection and use of PI emerge in multiple settings across the assessment timeline. A starting point is almost always when the test taker registers or signs up for a testing event. Whether that action occurs online or in-person, PI will be collected. Another point is where test taker PI may be involved is when the test administration is arranged and/or conducted, including if services associated with the test administration in involved (e.g., remote proctoring). A final point is whether and how the PI is used to determine the test outcomes (e.g., will administration or scoring decisions be taker). A related question in test administration is whether any sensitive or specially protected PI is used (e.g., collection of medical information to determine appropriate accommodations or use of a biometric identifier to determine outcomes).

A further issue emerges under international privacy laws that do not view test results in a consistent manner. In general, PI is defined as information collected from an individual, so raw answers to test questions might well be considered as "personal;" however, jurisdictions vary as to whether test outcomes are deemed to be personal. Thus, a testing organization needs to address this issue in its policies and procedures. Many sound business practices related to data privacy derive from European laws in this area, including the General Data Protection Regulation (GDPR), and other national privacy laws that generally follow the GDPR (e.g., Brazil, India).

While legal requirements regarding PI continue to evolve around the globe, common principles have emerged that are generally reflected in most of those requirements, for example, the OECD Privacy Principles (2013) and the NIST Privacy Framework (SP 800-53, revision 5, 2020).

Given the various international laws, and emerging principles, a general outline a testing organization should follow for privacy compliance involves at least these eight steps:

(1) conducting an inventory (or mapping) of what PI is collected, used, shared/transferred, and stored;
(2) determining the lawful basis for any data collection, use, processing, storage, or any transfers of data based on applicable legal requirements;
(3) developing a written privacy policy addressing the types of PI, uses of PI and organizational purposes, security of PI, and any disclosures of PI;
(4) determining what notices are required to give to individuals about the testing organization's PI procedures, what consents are required to obtain from individuals, and what rights an individual has under applicable privacy laws and regulations;
(5) conducting a privacy impact assessment (PIA) or similar risk assessment to document how the organization implements privacy principles and balances the need for protecting individuals'

privacy against other needs (e.g., administering a test fairly for all test takers, protecting the organization's IP);

(6) reviewing all third-party agreements with vendors/suppliers (e.g., test developers, test administration providers, scoring service providers, remote proctoring providers, cloud hosting providers) with whom the organization shares test takers' PI (or those who share PI with the organization if it is only a processor of PI);

(7) referencing the organization's data security plans (see Chapter 8) to assure that PI is securely protected; and

(8) developing internal procedures for responding to requests from individuals under applicable privacy laws and regulations and for training staff regarding their responsibilities concerning PI.

This chapter sets out guidelines all TBA programs should address as a matter of good practice. Depending on the jurisdiction(s) an assessing program operates in, there may be additional requirements imposed by law. Often organizations will need to choose an appropriate and proportional balance between test security and test taker privacy. Readers of this chapter may find it useful to refer to the glossary for definitions of key terms (e.g., anonymization, data controller, data breach, data processor, personal data, processing, pseudonymization, sub-processor, etc.). In addition, a number of important data privacy documents are listed in the References section of these Guidelines. Readers are encouraged to keep abreast of changes, considering the evolving nature of this area.

The Association of Test Publishers published a comprehensive guide for testing organizations (2017) which focused on practices required for the GDPR. The Association of Test Publishers' International Privacy Subcommittee produced a series of concise and informative Privacy in Practice Bulletins (Association of Test Publishers, 2019-2022), which address different aspects of privacy impacting assessments (additional Bulletins may be published). ATP also produced *Privacy Guidance When Using Video in the Testing Industry* (ATP, 2020), which is particularly relevant for organizations implementing remote proctoring using video. These documents are useful references for those wishing to understand more about guidance on privacy with TBAs.

Guidelines for Privacy in Technology-Based Assessments

9.1 A testing organization needs to identify and follow the privacy laws and regulations that apply to it. To accomplish compliance with these laws and regulations and to inform all stakeholders, the organization needs to develop, adopt, and implement a written privacy policy that is transparent and easily understood by the relevant stakeholders.

Comments: It is important for a testing organization to identify and follow the applicable data privacy laws and regulations in all jurisdictions that apply to them and the intended participants to better protect test takers' PI and related test data. Where a testing organization operates in a jurisdiction that requires processing to have a lawful basis, it needs to ensure it has articulated a lawful basis for the processing, such as contract performance, legitimate interest, consent, or other ground under applicable law.

9.2 When a testing organization transfers personal data across national borders, it should follow any applicable requirements for the lawful transfer of such data.

Comments: Some jurisdictions prohibit or restrict the lawful transfer of personal data to other countries unless specified requirements are met (e.g., European GDPR). Different approaches are taken in other jurisdictions; assessing organizations should be aware of and follow these requirements, including any that extend to further onward transfers.

9.3 Where a testing organization use biometrics (e.g., facial recognition, palm-vein scans, iris scans) to identify individuals or detect possible testing irregularities, it should be aware that many jurisdictions have laws relating to such use and needs to follow the requirements in such laws.

Comments: Biometrics covers a number of different technologies that are based on analyzing human physical characteristics, including fingerprints, iris scanning, voice recognition, vein or palm analysis, and facial recognition. Assessing organizations should determine which, if any, of the technologies they are using generates biometric data. This includes any use by vendors that provide assessment delivery services. In some jurisdictions, biometrics are considered sensitive data, which are deemed higher risk than other personal data and therefore require greater protection [see ISO/IEC 19784 standards as well as other technical subcommittees (e.g., SC17 and SC27)]. Specific to facial and voice recognition, it is important for a testing organization to distinguish between use for determining whether a person is someone of interest (e.g., identifying cheating behavior), confirming a person is the same as the one who previously provided an identification, or just detecting if multiple faces appear on a computer screen during a testing session.

9.4 A testing organization needs to identify and appropriately categorize its role, and the roles of all entities involved in delivering the program, according to the requirements of applicable privacy laws, for example, as controller, processor, or sub-processor.

Comments: Privacy laws typically allocate responsibilities to organizations processing personal data based upon their given role. For example, the assessment sponsor might, under such laws, be designated as the "data controller," "covered business," or "responsible party," and its vendors may be "data processors," "service providers," or "operators." Because organizational responsibilities flow from the proper role accorded to an organization, getting this right is an important early and fundamental task. It is also important to note that just as an assessment sponsor may use multiple service providers within the assessment program, more than one entity may be considered the data controller in respect to the personal data collected and used as part of that assessment program.

9.5 A written agreement for processing personal data should be in place with each processor involved, including any service vendors, both directly between the controller and the processor and between the processor and any sub-processor (and any further levels down).

Comments: The trend among modern privacy laws is to require a written agreement between parties involved in processing personal data. Even if not required according to applicable laws, it is *strongly recommended to have such an agreement in place. The written agreement should specify a processor is required only to process personal data in compliance with the instructions of the controller and to delete or return all personal data at the end of the contract or when instructed to do so. If the processor is permitted to retain deidentified, aggregated, or anonymized data, such options may be specified. The agreement should require the processor to inform the controller if and how its instructions may not be in conformance with any applicable privacy laws of relevant jurisdictions and to request clarification of those instructions. Privacy laws may provide a specified definition of events that qualify as data breaches; an assessing organization should familiarize itself with any such definitions in laws that apply to it. Data breaches are often broadly defined to include unauthorized access to, loss, destruction, and alteration of personal data as well as disclosure of personal data (see 9.13). The agreement may also stipulate, according to the requirements of applicable laws, that the processor should inform the controller of all sub-processors that it utilizes and include a process for the controller to review and object to any changes to sub-processors. The written agreement should also include a requirement that the processor cooperate with the controller to respond to inquiries from applicable regulators, assist in conducting PIAs, and assist with responses to individual test takers. Finally, the written agreement should indicate all of the security requirements imposed by the controller on the processor (see Chapter 8).*

9.6 A testing organization should only collect the minimum personal data needed for the requirements of the assessment process and retain it only as long as needed for the stated purpose(s) for which it was collected or reasonably related purposes (as permitted by applicable laws). Access to personal data should be limited to only necessary personnel, including the use of legally-binding written disclosure agreements.

Comments: The testing organization should utilize the concept of data minimization – limiting the collection and use of personal data to only the information needed to operate the systems and/or provide the services. The processing of personal data should be limited to the use specified in the organization's privacy notices and policies. Any use of personal data for purposes other than conducting the assessment should be disclosed in advance to the test taker. If the testing organization adds a new purpose(s) for collecting/using PI, it is obligated to provide a new, updated notice of such purpose(s) to test takers. Consistent with the security principle of least privilege, employee/contractor access to PI should only be afforded to those individuals that require it for purposes of discharging their duties relevant to the administration of the testing program. As stated, the testing organization should adopt and follow written policies and procedures to ensure that limited access to PI is observed in practice.

9.7 A record of processing activities should be maintained, reviewed, and updated at least annually, or whenever the organization changes its processes/data flows. This record should include a data inventory (i.e., a mapping showing where personal data are found in its systems), how it is used, and to whom it is disclosed. The record also should state the purposes of the processing, the categories of individuals and personal data processed, cross-border data transfer details, retention requirements, and security protections.

Comments: This inventory/mapping of PI also will enable the testing organization to locate all relevant PI in order to respond to a request from a test taker. The need to respond quickly to such requests means that the organization will benefit from using the mapping/inventory, along with an automated system for generating responses. Many organizations have experienced significant costs to deal with requests under applicable privacy laws.

9.8 Where practical, personal data captured during the assessment process should be stored and transmitted in an encrypted and/or pseudonymized form to reduce the risk of unauthorized access or disclosure of personal data.

Comments: Use of PI with such protection decreases the risk of loss or unauthorized access to PI (see also guideline 9.13). Although pseudonymized personal data may still be considered personal data, pseudonymization may be considered an appropriate technical security measure, be relevant with respect to any risk assessment required to transfer personal data across borders, and also mean that what would otherwise be a reportable data breach may not, in fact, require reporting. Encryption is also very helpful in removing privacy risks. When using pseudonymization, a testing organization needs to be able to demonstrate it does not possess the decryption key or key to re-identify the person whose personal data was pseudonymized. Another method of protecting PI is by anonymizing the data and aggregating for uses (e.g., research, norming), to ensure specific individual PI cannot be re-identified. Generally, anonymization requires that PI cannot be re-identified,

9.9 The retention period for the different types of personal data processed by the assessing organization should be documented.

Comments: A testing organization should retain PI only as long as it is needed for the purpose(s) for which it was collected or for such period of time as is reasonably related to those purpose(s), as permitted by applicable laws. The period of retention may vary with different types of data and generally depends on the sensitivity of the data. For example, it is common to keep copies of government identification cards or biometric data used for identification for a short period but to maintain assessment scores and pass/fail records for a longer period, especially if the test taker has the right to an appeal/challenge of test scores/reports.

9.10 When the retention period has passed, or if there is no longer a need to retain data, personal data should be securely deleted according to established industry standards in such a way that it cannot be reconstituted.

Comments: As part of its retention policy, a testing organization should identify a data deletion policy to define which data is deleted within specific periods of time or based upon agreed criteria (if it is not possible to define a precise time period to apply in all cases).

9.11 A testing organization needs to inform all test takers in a clear and easily accessible privacy notice and privacy policy about what personal data are collected, how the data are used, and whom to contact if they have any questions.

Comments: A testing organization should publish a privacy notice setting out its privacy policies so test takers and other stakeholders know how the organization will treat personal data. This privacy policy could be posted on the organization's website or provided as a link as part of a test taker agreement presented during test registration and/or at the start of the test event. In addition to a privacy policy, most privacy laws and regulations require organizations that collect and use PI to provide appropriate notice to every individual about specific actions that may be taken (e.g., the use of automated decision-making, artificial intelligence [AI], video surveillance, or biometrics). Test takers should be informed in advance of taking the assessment what types of personal data are captured during the assessment process, who is responsible for the management of the personal data (including who is the controller), what tools are used to collect personal data (video, audio, biometrics, AI, locked-down browser, key strokes, etc.), what personal data are collected from the tools, how long the data are retained, how the data are secured and protected, how the assessment results will be used, and to whom the assessing organization gives access to assessment results and other personal data.

9.12 Test takers need to be informed in advance of taking the assessment regarding their rights with respect to accessing their data.

Comments: Unless there is a lawful basis otherwise, test takers should have a right to see personal data held on them within a reasonable time period. The controller should define in written policy rules whether test takers can request to see, correct, or delete their personal data and what criteria are used to consider such requests. The assessing organization should have a defined policy for dealing with requests that includes telling data subjects how they can exercise their rights under applicable laws, such as by contacting a dedicated email reflector set up for this purpose. The contract between the data controller and any data processor, such as a testing services provider, should also specify how the vendor will inform the data controller if any test takers make requests directly to the service provider. Unless otherwise agreed between the data controller and data processor, it is the data controller's responsibility to address privacy requests from test takers, and the data processor shall assist in accordance with the data controller's instructions.

9.13 In the event of a data breach where personal data have been improperly accessed, and test takers may be identified (e.g., the data are not securely encrypted or pseudonymized), the testing organization must comply with applicable breach notification laws/regulations.

Comments: Each assessing organization should have a plan for addressing such situations, starting with a security incident that needs to be investigated to determine if it is a data breach (see Chapter 8). In the event that a situation as described does occur, the plan should be followed to confirm if a data breach has occurred and then ensure the appropriate actions are taken. Some breach notification laws and regulations are part of a general privacy law; others are standalone *provisions. Some laws/regulations require the entity suffering the breach to notify only the regulator; others require notice to the affected individuals/test takers. Privacy laws may include a specified definition of events that qualify as data breaches, so testing organizations should familiarize themselves with any such definitions in laws that apply to them. Data breaches are often broadly defined to include unauthorized access to, loss, destruction, and alteration of personal data, as well as disclosure of personal data. Additionally, the time period and requirements for giving notice in the event the testing organization confirms a data breach varies by jurisdiction. Because of these variations, a testing organization needs to research and understand which law(s) apply. A more comprehensive discussion of the issues related to data breaches is found in Chapter 8 (Security).*

9.14 A testing organization's use of automated decision-making (software that merely automates a manual process such as scoring or test assembly) within the assessment process should be performed in a way that is ethical and fair, respects individual rights, and complies with applicable laws.

Comments: Prior to implementing any automated decision-making within the assessment process, the testing organization should conduct a thorough review of its solutions and procedures to ensure they result in fair treatment of diverse populations and have a demonstrated track record of operating without bias or discrimination. An assessing organization should follow the applicable rules with respect to the use of automated decision-making as regards test taker privacy rights, including providing transparency with respect to personal data processing involved and ensuring appropriate human involvement, as well as potentially providing test takers with assessment alternatives that do not involve the use of automated decision-making, which may be required under some national privacy laws/regulations.

While global AI regulations are generally not in effect at the time these guidelines are being prepared, some controversy exists about whether automated decision-making should be considered AI (see Section) It is important to distinguish between AI and automated decision-making (i.e., software that merely automates a manual process, such as scoring or test assembly). However, some privacy laws/regulations (e.g., GDPR) have requirements around the use of automated decision-making; where the controller uses automated decision-making, a test taker has the right to be informed about its use and given a reasonable explanation of how the automated decision-making occurs. Note: Specific guidance on AI systems is found in 9.15 and 9.16, and further discussion of AI is provided in Part IV.

9.15 A testing organization's use of a specific type of AI system (e.g., machine learning, Natural Language Processing) should be subject to thorough and ongoing evaluation to assure ethical use, fairness and quality, especially to mitigate bias and discriminatory impacts.

Comments: Although regulation of AI is not in effect at the time of this writing, prior to implementing an AI system solution within the assessment process, a testing organization needs to conduct a thorough review of the system and process to analyze if the use of the system is fair to *test takers' privacy rights and complies with any applicable laws/regulations. The testing organization needs to document its findings because such information may be required by regulators to show that the AI system is providing fair treatment of diverse populations and operates without bias or discrimination. Care should be taken to appropriately source any data used to "train" the AI system; the output decisions or classifications resulting from the use of AI or algorithms should be reviewed and remediated if there are material discriminatory or biased impacts (see also Chapter 11). A similar analysis needs to take place after an AI system is in use, to document the system continues to be fair and unbiased. When using AI to score either items or tests, care must be taken to avoid promoting existing patterns of privilege or score bias.*[5] *A testing organization must take particular care around the use of AI that may impact children and use it only where necessary after conducting a PIA and balancing children's privacy interests against those of the program*.

9.16 A testing organization needs to modify its appeals procedure if it uses AI so test takers can challenge their scores to a human reviewer to determine whether a scoring decision was fair and appropriate.

Comments: Where AI is used to flag test taker behaviors or is used by testing organizations to make decisions about individuals, a human should always be the ultimate decision-maker, and the AI process should not be used alone to make significant decisions related to test takers (see European Commission, 2019, 2020). In addition to human involvement in the development and training of the AI system, a testing organization should document an appeals procedure that involves review by a human, so there is a check on the propriety of the decision resulting from the use of AI (see additional information in Part IV). Even if a human is involved in generating an AI system algorithm (i.e., a so-called "human-in-the-middle" AI), some types of AI algorithms have no human intervention involved in the outcome process. Thus, a testing organization should ensure test takers have an opportunity to have AI-generated results/decisions reviewed by a person upon request.

9.17 All employees, contractors, agents, or others involved in the organization's assessment process should exercise all reasonable efforts to ensure personal data are collected and processed in a safe and accurate manner. If an error or inaccuracy is identified, it should be addressed promptly.

Comments: Privacy laws tend to avoid prescribing specific measures. Rather, they expect organizations that process personal data to make a judgment regarding appropriate measures based on the personal data and processing activities involved, mindful of the risks involved and possible measures that may be employed. A testing organization should set appropriate measures

[5] See, for example, IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition. IEEE, 2019. https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/ autonomous-systems.html.

in advance, review these regularly, and ensure any data processors follow similar measures based on the nature of their activities.

9.18 Where personal data are inaccurate or errors are made in the processing of personal data, test takers should have the right to have errors rectified in a timely manner.

Comments: A testing organization needs to provide test takers with appropriate information on how they can request information under any applicable privacy law/regulation, including exercising the right to correct PI. To keep track of those test taker requests, the testing organization should develop and implement internal procedures for handling them to comply with requirements of applicable privacy laws/regulations

9.19 A testing organization should develop, implement, and maintain appropriate technological, physical, and organizational security measures that meet established industry standards designed to protect personal data from destruction, loss, alteration, and unauthorized disclosure, access, or processing. Such security requirements are found in most national privacy laws and regulations, so a testing organization needs to consider the specific requirements applicable to the jurisdictions in which it operates.

Comments: See Chapter 6. Data Management for further discussion on data security and protection and Chapter 8. Test Security.

9.20 A testing organization should engage a third party at least annually to evaluate its information security measures using established industry standards.

Comments: See also Chapter 6 (Data Management) and Chapter 8 (Security).

9.21 Legally reviewed confidentiality agreements need to be in place with all individuals who have access to personal data, including employees and contractors of a controller and its processor(s).

Comments: All individuals with access to personal data should, upon hiring and at least annually thereafter, receive training in their responsibilities relating to processing personal data, and their understanding following training should be assessed. Confidentiality terms may be free-standing or included within broader contracts, such as contracts of employment or vendor services agreements. Assessing organizations should also require that their vendors have written confidentiality provisions in place with their respective employees and contractors.

9.22 Appropriate enhanced safeguards should be implemented to protect data considered "sensitive," including biometric data for the purpose of uniquely identifying a person, health data, race/ethnicity data, or children's personal data.

Comments: Such data are considered higher risk in that if they are compromised in a data breach, the possible harm to the test taker would be greater as compared to less sensitive forms of personal *data. Assessing organizations should consider specific enhanced safeguards for sensitive data and document these as part of its broader technical and organizational security measures.*

10. FAIRNESS AND ACCESSIBILITY

Background

The AERA et al. (2014) *Standards* define fairness in testing by stating a test is fair if it "…reflects the same construct(s) for all test takers, and scores from it have the same meaning for all individuals in the intended population; a fair test does not advantage or disadvantage some individuals because of characteristics irrelevant to the intended construct" (p. 50). Of course, fairness in testing extends beyond the test itself to all aspects of the testing process, including test development, administration, scoring, and score reporting. For this reason, the AERA et al. (2014) *Standards* describe four perspectives on fairness, which are fairness (a) in treatment during the testing process, (b) as lack of measurement bias, (c) in access to the construct(s) measured, and (d) as validity of individual test score interpretations for the intended uses. These perspectives are reflected in the guidelines provided in this chapter.

Ensuring all test takers have sufficient "access" to the test means test takers are able to demonstrate their proficiencies without being hindered by construct-unrelated characteristics of the testing process. In this chapter, we first discuss issues of access and then turn to general issues of fairness in testing. Following a brief discussion of these issues, we present guidelines for accessibility and fairness in technology-based assessment (TBA).

Accessibility

The adoption of digitally delivered tests has expanded opportunities to increase the accessibility of test items and interfaces. Whereas the concept of accessibility was once equated with test accommodations for students with disabilities, accessibility is now a concern for all test takers and includes universal design and accessibility supports (see Chapter 1, and Lee et al., 2021). Accommodations have traditionally been treated as changes to test conditions designed to increase test takers' access to the test for specific subgroups identified with a disability, as well as for multilingual learners (International Test Commission, 2018a). Accessibility, however, is treated as an integral component of all phases of the test development process and aims to eliminate barriers to measure a targeted construct in a valid and reliable manner. Today, the aim of accessibility is to minimize construct-irrelevant variance ("CIV") and maximize construct relevance for all test takers.

One way to promote accessibility is to permit supports and accommodations for individuals likely to encounter construct-irrelevant barriers during testing, including, but not limited to, elderly test takers with less familiarity with technology, those with disabilities or second language learners. Accommodations target a need associated with a specific disability by altering the manner in which test content is presented to a test taker or by altering the tools a test taker uses to navigate and respond to test questions. Accessibility during test administration falls into four broad categories: accessing content, interacting with content, response production, and interface navigation. The first three categories are item-specific and address the three phases test takers pass through as they engage with a test item. The final category applies to the test delivery system more broadly and focuses on the way in which a test taker employs various functionality built into the delivery platform to engage with items and the test as a whole. It is important to note accessibility is also a concern for educators and test administrators who may interact with a testing platform to register, assign accessibility settings, and access reports.

All tests require test takers to engage with test content. In many cases, this engagement occurs cognitively as the test takers think about and work on the problem presented. In other cases, the content may be embedded in a clinical assessment. In digital environments, however, interactions increasingly require test takers to engage with digital representations of content. As an example, some science items require test takers to manipulate digital tools to simulate an experiment. Similarly, some mathematics items allow test takers to engage with digital models and tools as they work through a problem, as do many credentialing exams. In addition, some items that measure social science skills require test takers to work with content presented in different texts, images, videos, and/or sound files.

Some digitally delivered items require test takers to navigate and manipulate content using a mouse, track pad, or finger taps and drags. For test takers who have challenges or limited familiarity using these devices, interactions with digital content interfere with their application of the targeted construct. Designing the interface and test delivery system to support keyboard (tab/enter/arrow) navigation allows test takers to use alternate communication devices (assistive technology [“AT”]) to navigate and interact with digital content. Online systems need to allow for other keyboard shortcut inputs for test takers using screen readers (and refreshable braille devices) to facilitate reading, review, input, and navigation of the various interfaces.

For test takers who experience challenges working among and processing multiple pieces of content tools that selectively mask content, it may be helpful to scaffold (support) interactions with content or present auditory background stimulation to help them focus and interact with the various components of item content. In addition, extended timing and breaks may be required by test takers with information processing needs and/or due to the increased time required to navigate and interact with test content using accessibility supports and/or assistive technologies.

Test items require test takers to produce a response. Similar to the interaction of content, response production requires test takers to employ one or more devices to input and/or manipulate content to create a scorable response/product. The same challenges and accompanying solutions associated with interacting with content are applicable to response production and interface navigation. For test takers who experience difficulty using a keyboard, speech-to-text or use of alternate keyboards allow them to produce text-based responses to open-response items. Some test takers may require a scribe to input responses. Finally, intuitively designed interfaces that limit the cognitive demands required to navigate among items and make use of interface options may minimize the effect of construct-irrelevant factors that can influence test performance. Interactions that accept only mouse input should be avoided unless they are necessary to provide a response for the construct being measured (e.g., a drawing supplied for a drawing assessment). If the interaction includes ”drag-and-drop” behavior, the interaction must have keyboard access that allows users to select and move the selectable objects into targets and include the ability to remove selectable objects from targets. The test taker should be able to understand the current association of objects through text-based information (typically provided by visually hidden information only available to AT).

Merely providing a "technically accessible" solution may not be adequate in an assessment context because users of AT often need to keep more pieces of information in their short-term memory. Tasks that require a large amount of information may begin to strain the memory ability of the candidate in addition to the construct intended to be measured. This can be especially true of test takers for whom the language of the test is not also their native language.

Technology-enhanced items (TEIs) may also provide challenges to individuals with disabilities. While TEIs can be made to be technically accessible, cognitive overload and memory requirements (e.g., navigating back and forth between item components) can be excessive and may therefore be inappropriate for certain users of AT and users of other accommodations.

Interoperability Accessibility Standards. The Web Content Accessibility Guidelines (“WCAG”), developed and published by the W3C, promote common standards of web accessibility throughout the world.[6] These guidelines not only provide guidance for special accessibility needs but detail success criteria that increase access for all users. When these standards are followed, test developers will less likely need to make modifications to their applications to meet the needs of a specific AT device. Technically meeting all success criteria within WCAG does not guarantee good access to assessment applications. There are many areas that require judgment calls as to the best organization of information and description of images. Content authors and editors must take responsibility for the alternatives provided for AT. Furthermore, there is often additional development work required to meet specific environments, like web browsers and screen readers, to make applications usable for people with accessibility needs.

In the early 2000s, the Question and Test Interoperability (“QTI”) standards were introduced to support the exchange of test content across item development and test delivery platforms. In 2008, an extension to the QTI standards was developed that aimed to develop a standard approach to embedding accessibility supports in item content. Initially termed the Accessible Portable Item Protocol (“APIP”) Standard, this set of specifications focused on three components of an accessible test design and delivery. The first component pertains to specifying supplemental and alternate representation of content that allows the item to be accessible to test takers with various access needs (e.g., braille reader, alternate language, text-to-speech, American Sign Language, definitions of key terms, etc.). The second component focused on accessibility supports embedded in a test delivery itself (e.g., magnification, alternate contrast, content masking, etc.). The third and perhaps most novel component provides a standard mechanism for specifying the access needs for each individual test taker, termed Personal Needs and Preferences (“PNP”). The PNP serves as a control center that specifies to a test delivery system which components of item content are presented to the test taker and which access

[6] https://www.w3.org/WAI/standards-guidelines/wcag/

tools embedded in the test delivery system should be available and/or activated for the test taker.[7] PNPs can also contain information about the test taker's session environment, such as special equipment, special system settings (e.g., enlarged cursor), room settings, medical requirements, non-computer supports, or any other concern.

Recently, QTI and the APIP standard have been refined and integrated into a single, updated standard termed QTI 3.[8] Adherence to the QTI 3 standard is an effective mechanism for supporting the accessibility of digitally delivered tests. Programs taking innovative approaches that introduce new features and supports to their programs can use QTI 3 (or a current release subsequent to this writing), which allows for extension points, and programs can add their enhancements in a more predictable (and interoperable) exchange format.

The Student Interoperability Framework enables state/district codes to be included within the Test Accommodation structure. While states (or testing programs in general) may have their own codes, there are other multi-state consortium standards, such as Smarter Balanced, that employ a standard set of accommodation codes. These codes may or may not be usable by various assessment delivery platforms.

Equity Issues in Technology-Based Assessment

Equity in assessment may be defined broadly as the development and maintenance of structures and/or systems designed to provide test takers with the tools they need to access the assessment and demonstrate their knowledge/understanding of the content through the assessment.[9] When referring to TBA specifically, an equity focus requires attention to additional barriers created/made possible by how the assessment is administered. The extent and influence of these barriers depend on the administration context (e.g., familiarity with digital interface for in-class TBAs or Internet access in rural, low-income communities for at-home TBAs). When addressing equity issues in any assessment, it is critical to consider the roles of *power*, *authority*, *access,* and *privilege* at every stage of the assessment process, including assessment development, administration, and scoring (Randall, 2021). When considering TBA, equity issues related to *privacy* become similarly critical. In any case, we maintain that just as socio-cultural and socio-political contexts are constantly shifting, evolving, and emerging, so must our conception of issues related to equity, diversity, and inclusion in technology-based assessment.

Structural biases built into the information organization system can serve to silence certain segments of society (Heffernan, 2020). Developers of TBA must attend to issues of power or, more importantly, *disempowerment* by critically considering who holds power and including those who, even if unintentionally, are disenfranchised by the assessment. In particular, the impact of the assessment on

[7] See https://www.imsglobal.org/sites/default/files/spec/afa/3p0/information_model/imsafa3p0pnp_v1p0_InfoModel.html

[8] https://www.imsglobal.org/spec/qti/v3p0/guide

[9] This definition is consistent with extant definitions in education. For example, see National Academies of Science, Engineering, and Medicine (2019), which defines equity as including unequal distribution of goods and services based on need.

traditionally minoritized populations (e.g., ethnic /racial groups, low socioeconomic status, physically disabled) should be thoroughly considered. For test takers from the most marginalized populations (e.g., immigrants, immigrants of color, individuals with housing insecurity) with limited social or political power, the potential, or perceived, negative consequences of TBA can be profound.

Context matters, in particular cultural context, which is important across and within countries. Culture is always operating and must be the lens through which we view educational practices (Battista, Ellenwood, Gregory, & Higgins, 2015). For example, the introduction and effective use of technology in disadvantaged populations (e.g., geographic/rural, economic, technological) may require accompanying the technology use with a complete change in teaching practices (Powers, Musgrove, & Nichols, 2020), along with grounding the instruction in the student's lived experience (Azano & Stewart, 2015). Using assessments across cultural contexts involves its own set of difficulties (Greenfield, 1997; International Test Commission, 2018b); one could only surmise that using TBA across contexts would be even more difficult.

In addition to attending to potentially negative outcomes stemming from differential socio-political power, test developers must also consider other forms of power (e.g., financial, geopolitical) that may limit test taker access to TBA. Indeed, a high-quality home testing experience requires access to adequate testing facilities and equipment. Test takers must have the requisite technological knowledge to use the equipment. Such knowledge is closely tied to socioeconomic status (Ercikan, Asil, & Grover, 2018). Test takers must have use of an electronic device with Internet access and the bandwidth to support that device during testing. Moore et al. (2018) found among students taking the ACT college admissions test that 14% had access to only one electronic device in their homes, and 56% of those students reported the device was simply a smartphone. Moreover, approximately 15% indicated their home Internet service was unpredictable or terrible, which poses an additional problem related to access as presentation and scoring methods reliant on high digital transmission speed can adversely affect those without high-speed Internet. The COVID-19 pandemic has brought disparities in connectivity into sharp focus (Herold, 2020). Even for test takers who can work around issues related to access to the necessary technological tools for assessment, online/remote assessment proctoring often requires additional financial resources often passed along to the consumer. The overall effect is technology use exacerbates the disparities in achievement across socioeconomic status (Chiao & Chiu, 2018; OECD, 2021). Traveling to a test center may also involve a differential burden across different groups of test takers, and the needs of elderly test takers should be considered in this regard.

Assessment developers must consider which groups, if any, are advantaged by the act of taking a TBA and engage in a critical process of mitigating that privilege. For example, algorithmic AI monitoring/ surveillance (i.e., surveillance that is performed by technology with the use of AI algorithms) is often based on flawed assumptions about "normal" (Howard & Borenstein, 2018; Mayson, 2019), which may disproportionately affect groups that do not fit developers' assumptions about ability, culture, race, or gender expression. As Swauger (2020) explained, cisgender, able-bodied, neurotypical white men are privileged as their movements/bodies will generally be categorized as "safe" and "non-threatening," whereas test takers from historically marginalized groups may have a very different experience. Swauger described the experiences of Black and Brown test takers being asked to shine more light on themselves when verifying their identities for a test or being unable to begin a test at all because the AI software used by proctor could not detect their faces--issues their white peers did not have to manage. Such examples of whiteness (or other dominant groups outside the USA) being privileged in TBA are not uncommon. Moreover, test takers who express their gender in ways that are not cis/heteronormative may experience similar obstacles when sitting for TBAs. Consequently, because test takers from these marginalized groups are aware they are vulnerable to these algorithmic misinterpretations, issues related to increased anxiety can contribute to poor test performance, thereby further privileging majoritized groups at the very real expense of minoritized test takers (see also Chapter 11, *Global Testing Considerations*).

Finally, in addition to considerations of power, access, and privilege, TBA developers are tasked with addressing concerns related to privacy (see Langenfield, 2020 for examples, as well as Chapter 9 of these *Guidelines*). Issues of equity with respect to privacy are particularly salient with respect to TBAs that rely on remote proctoring or AI monitoring. We refer the reader to the privacy chapter (Chapter 9) for a more detailed explication of issues related to privacy. Here, we focus on privacy issues as they relate to equity only. Such equity issues are of particular concern when technology is used to administer at-home assessments. For example, to engage with the assessment, at-home test takers must often agree to allow proctors access to their homes via the camera on the electronic device (consider test takers who live in crowded or chaotic conditions), which could increase test anxiety levels, thereby diminishing test performance (Flaherty, 2020). In an extreme case, Katz and Gonzalez (2015) found many immigrant families feared state surveillance through school-issued laptops. This fear could effectively preclude the use of technology at all. When employing the use of technology to monitor test taker behavior/ environment during at-home assessments, test developers must not assume all test takers live in stable, well-lit, pristine homes they would be proud to display/share with strangers. Indeed, the emotional impact--due to shame, fear, uncertainty--should be considered when making decisions related to monitoring and ensuring test taker privacy.

Guidelines for Fairness and Accessibility

Guidelines for Accessibility

10.1 A testing organization should make all aspects of the assessment lifecycle as accessible as possible, including test information, test registration, accessibility and accommodation request forms, login screens, assessment interfaces (including sample tests), and test results. This will allow personal accessibility for all test takers, including those with accessibility needs.

10.2 A testing organization should establish clear guidelines regarding which accessibility and accommodation supports are available to all test takers and which must be specified in advance for select test takers.

Comments: Testing organizations should establish where it may be appropriate to allow the assignment of accommodations without prior approval (e.g., highlighter to highlight text, tiered accessibility frameworks).

10.3 A testing organization should establish clear definitions of the construct(s) targeted by the test and specify which categories of accessibility may or may not change the construct measured.

Comments: Testing organizations should establish clear criteria regarding when the provision of an accessibility support alters the assessed construct in a manner that voids the intended inference based on the test score (Abedi & Ewers, 2013).

10.4 A testing organization should require WCAG compliance.

Comments: Testing organizations, including authoring platforms and other internet accessibility features, should strive to be WCAG compliant (currently WCAG 2.1 Level AA). for item authoring and rendering, as well as all aspects of the interaction with test takers. People with different access needs are critical to evaluating the appropriateness of alternatives provided to test takers or restrictions placed on them.

10.5 A testing organization should allow breaks, extended testing time, and other accessibility supports and accommodations to standard test administration conditions subject to valid measurement of the construct and test taker needs.

Comments: Accessibility support and accommodations should be provided wherever possible, but if they may change the construct measured by the test and reduce the validity of score interpretations and uses, research should be conducted to ascertain the effect of the supports and accommodations on score interpretation and use. Consideration and implementation of supports and accommodations must consider both validity of score interpretation from both construct representation (changing the construct measured) and CIV (e.g., providing access to demonstrating the construct). See Abedi and Ewers (2013), AERA et al. (2014), and Sireci and O'Riordan (2020) for discussions of these issues. AT may require more time to navigate computer-based content, and promote testing fatigue, which would suggest breaks may be needed. Thus, the use of supports and accommodations may require extended test administration time.

10.6 Test authoring platforms should allow authors to input accessible alternatives to content.

Comments: Examples include text-based descriptions of images, or indications images are "decorative;" long descriptions (where appropriate); alternatives for time-based media (e.g., captions, audio description of video, transcripts); appropriate HTML and Accessible Rich Internet Applications annotation applied to tables, figures, and navigation elements of item content to support text-to-speech presentation of item content. The authoring platforms should include the types of supports and accommodations available for the item type the author is working on, so authors are more likely to develop items that fit these presentation/response mechanisms.

10.7 A testing organization should comply with standards that facilitate the implementation of interoperable accessibility features, such as the current version of the QTI standard.

Comments: Item authoring should employ appropriate QTI encoding to specify alternate representations of content, including Braille representations, signed language representations, alternate language representations, text-to-speech markup for pronunciation, simplified language versions of item content (partial or complete), and key word definitions.

10.8 Test taker registration systems should employ appropriate tags to document the access needs of individual test takers, such as in a PNP profile.

Comments: Test delivery platforms should support embedded accessibility supports, including magnification of item content that allows for text reflow to prevent horizontal scrolling, enlargement of text that allows for text reflow (with exceptions for specially formatted content like poems), alternate display of text and background colors, text-to-speech and/or audio representations of item content, and masking of item content. If methods other than PNPs are used to document test takers' accessibility needs, they should be validated, transparent, and transferrable across systems.

10.9 Test delivery platforms should be QTI (or other consensus industry standard) compliant to activate appropriate embedded accessibility supports for each individual test taker and present appropriate alternate representations of item content for each individual test taker.

10.10 Test delivery platforms should interact with AT devices such as alternate keyboards, single and dual switch mechanisms, speech-to-text software, text-to-speech software, and refreshable braille displays.

Comments: It is important to allow access to the test delivery platform prior to testing to confirm the functionality of AT devices (e.g., practice or sample tests, accessibility reviews).

10.11 Tablet-based test delivery platforms should design interactivity to function appropriately for test takers with fine-motor skill needs and fingertips of various sizes such that accuracy of responses is not adversely impacted.

Comments: Interactivity may also involve the use of gestures for test takers to navigate their tablets and screen readers to describe their actions.

10.12 Testing programs should establish clear guidelines regarding when and whether test content will be presented in a paper-based form for test takers who cannot be supported appropriately in the digital test delivery platform.

Comments: This includes test takers who require embossed braille and tactile graphics.

10.13 TBAs should avoid input that requires a second mouse button click.

Comments: Test takers with some motor disabilities will be unable to double-click with ease, so such actions will increase response time and stress.

10.14 TBAs should use vector-based graphics over pixel-based graphics where possible.

Comments: Vector-based formats such as scalable vector graphics *allow for scaling at high resolution. When including pixel-based content (i.e., photographs), use over-sampled versions that display at half the width and height. This allows for magnification at 200% without pixelation (jagged/blurry images).*

Guidelines for Equity in Technology-Based Assessment

10.15 TBAs should be designed within a framework of diversity and inclusion.

Comments: Assessments should be evaluated at every stage of development to ensure that cultural language, practices, and experiences centered in the dominant culture (e.g., whiteness) do not form the basis for the assessment. The assessment should be designed to be appropriate for all groups of intended test takers. Ideally, the development team for TBAs would include specialists from all major groups (e.g., racial/ethnic, gender, linguistic minorities) targeted by the assessment. Test developers should understand the challenges these populations face and be committed to representing the needs of these populations (especially historically marginalized populations) so all test takers can see themselves represented in the assessment. Specific attention should be given to intersectionality across groups in the intended testing population. Statistical analyses (e.g., differential item functioning) should be used in evaluating the appropriateness of test and item design for all test takers.

10.16 All test takers should be given sufficient time to become familiar with the testing environment prior to testing

Comments: All test takers should have an opportunity to use the testing equipment (actual or similar) prior to the actual examination (ideally during instruction) and should have access to practice material in the same format as the actual assessment prior to testing. Test takers should also have access to all approved supports and accommodations allowed on the actual assessment to verify functionality prior to testing.

10.17 When TBAs use remote monitoring, test takers should have the opportunity to become comfortable with the virtual proctoring software and environment in advance of the assessment.

Comments: Students should have the opportunity to interact with monitoring/surveillance technology prior to the actual test-taking experience, sufficient to satisfy them that they will not be *disadvantaged by the technology. If the interactions cannot be made satisfactory, a vehicle for the test taker to raise objections should be made available.*

10.18 Test administrators should ensure all test takers have equipment and connectivity that allow the proper delivery of the assessment (i.e., without adversely affecting timing or performance on the assessment).

Comments: All test takers should have access to an adequate environment (equipment, connectivity, surroundings, etc.) for taking the assessment.

10.19 Test administrators should ensure different groups of test takers are not differentially affected by technical disruptions that could adversely affect their performance.

10.20 When observing and video recording individuals at home or in a testing center, clear statements must be made with respect to how the personal privacy and data of test takers will be maintained.

Comments: Test developers/administrators should disclose all uses of test takers' personal data for any purpose other than that directly needed to deliver, score and report on the test. Any use for research purposes should be disclosed in advance and be in accordance with applicable privacy legislation in the jurisdiction. Test taker data should not be used for marketing purposes without the test taker's express written consent. Moreover, refusal to provide consent should not result in a penalty of any kind, and likewise, providing consent should not result in any direct benefit to the test taker.

10.21 In the case of remote proctoring, all unintended, negative consequences of monitoring should be investigated and removed or minimized.

Comments: It is good practice to consider all characteristics of test takers that could affect proctoring (remote and in-person), such as cultural and religious contexts in (e.g., respect head coverings where culturally/religiously required and to allow a female test taker to request a female proctor). Consider the use of test score verification tools as an alternative to "live" test monitoring/proctoring tools if such tools prove to be invasive or lead to increased test anxiety

10.22 When possible, the pool of remote (and in-person) human proctors should match the test-taking population in terms of demographics.

Comments: Human proctors should be trained extensively to reduce the likelihood of discrimination against test takers based on their bodies, identity, appearance, atypical movements, and/or race or ethnicity.

10.23 Algorithmic proctoring mechanisms should be extensively tested to ensure they behave accurately and identically across various groups of test takers.

Comments: The mechanism should be carefully pilot tested with test takers from various groups to ascertain any negative impact monitoring may have on test taker performance.

10.24 Automated scoring engines should be calibrated with all groups of test takers in mind.

Comments: As appropriate, automated scoring engines should be trained using representative test takers from all groups in the testing population. Care should be taken to ensure that automated scoring engines do not perpetuate or exacerbate human scoring biases and the validity of the scoring is consistent across groups of test takers. Identifying representative test takers should include training automated scoring engines using a range of assistive technologies (i.e., screen readers, text-to-speech, refreshable braille displays, alternative keyboards, and mouse emulators, such as switches) on varied platforms and operating systems.

10.25 Automated scoring engines should be monitored to ensure they record the same scores for equivalent responses across all demographic groups in the testing population.

11. GLOBAL TESTING CONSIDERATIONS

Background

Technology-based Assessments ("TBAs") are administered worldwide. Some TBAs are administered across multiple countries, and some are national or even local. Considerations in this global environment include translating and adapting tests for use across multiple languages and cultures, technology availability resources, and preparing test takers for the assessment experience across a wide variety of environments that vary with respect to technological resources. In this chapter, we first discuss these issues and then present guidelines in each area.

Translation and Adaptation

Translating or adapting an assessment should always aim to obtain a test form in the target language that (a) measures the same construct; (b) is fair and unbiased; (c) has sufficient reliability; and (d) is valid for its intended purpose. In some contexts, such as comparative research or a competitive pre-hire test in multiple languages, an additional aim will be to obtain scores from each language version that are comparable to those obtained by the test form in the source language. The International Test Commission *Guidelines for Translating and Adapting Tests* (2018), the AERA et al. (2014) *Standards* (2014); and references such as Dept et al. (2017), Grisay (2003), Hambleton et al. (2005), Hambleton and Zenisky (2011), Harkness (2003), and Iliescu (2017); provide robust guidance and references on translation and adaptation of tests in general. There is also a body of literature that documents how and why different forms of adaptation to local context and usage affect measurement (e.g., Allalouf et al., 1999; Allalouf & Hanani, 2010; Ercikan, 1998, 2002; Sireci, 1997; Sireci et al., 2005). In this chapter, we focus on guidelines specific to translating or adapting assessments in a technology-rich environment.

There is a substantial difference between situations where test developers are designing a new assessment and intend to make it applicable for two or more languages or cultures; and situations where an existing assessment that is already deployed in one language needs to be translated/adapted into other languages. For the first type of situation, it is essential to note good practice is to embed translation/adaptation in the test design and development process from the outset rather than to view it as a standalone component. This aspect has become even more salient in TBAs, where translatability, cultural appropriateness, and portability should be dealt with before the authoring stage. For the second type of situation, where an existing test needs to be translated, one must be prepared to consider various revisions to the existing language version of the test to make it a suitable source version, or starting point, for the translated forms of the test. The guidelines in this chapter endeavor to cover both situations.

In the preliminary stages of test design, the test developer determines the purpose of the test and the construct or domain being measured. That is also a good time to consider possible target populations, languages in which the items might be translated, and to investigate cultural differences between segments of the extended target population (ITC, 2018, 2019). It is recommended to seek advice both on portability of the construct or domain and on possible adjustments to item formats. Early definition of a multistage translation process and a pilot test of translated versions will significantly contribute to a controlled progression toward fairness, validity, and reliability in the different target versions of the test.

In case of a transition from one delivery mode to another, experience in international large-scale assessments has shown that a computer-based environment is more than just a different medium in which previously successful translation procedures can be applied. When setting up a new translation design, it is recommended to take into account the limitations and requirements as well as the range of opportunities offered by the technology used in the new testing ecosystem.

In all cases, for translation and adaptation of a TBA, it is advisable to put in place the following:

- documentation on the measurement characteristics of the measurement instruments;
- an agreed translation and quality assurance (QA) plan, including a format that translators can use;
- interoperability and translation data exchange standards. The most widely used translation data exchange standards are XML vocabularies, which means the vocabularies can be validated by means of other XML utilities (about XML standards in translation technology, see Roturier, 2019);
- a process to preview the source version and the target version of the assessment, preferably at any stage in the process; and
- an agreed process for revalidation of the new TBA.

Availability of Technology Resources

Despite the advancements in TBA and remote assessment, some populations (e.g., developing nations, rural communities in developed nations, etc.) may be inadvertently disadvantaged due to local infrastructure. When this occurs, there are means to use TBAs to support test taker needs. The goal is to offer a fair test for all test takers, and, as much as possible, the delivery of the test provides a comparable testing environment for all candidates, regardless of modality.

There are a myriad of ways in which exams are delivered using technology outside of traditional brick-and-mortar test centers. For example, large-scale delivery vendors offer as-needed testing to targeted populations to supplement capacity needs and to support rural markets of high need. Additionally, there are models in which exams are administered at training facilities, conferences, and in kiosks in retail locations. Each test delivery model may be problematic where technology resources are scarce. Among the models, there are several variables that can cause technical issues that impact test events, primarily Internet availability, connectivity, and stability. Depending on the testing modality and model, it may be important that there is a strong, uninterrupted Internet connection available throughout the testing event if the exam content is to be sent to the test taker's location in real time while the person is testing.

While it is strongly recommended that a system check be conducted prior to the test taker receiving any content to evaluate risks such as poor Internet connections or outdated security patches, it is important to remember any check will only provide a point-in-time snapshot. Internet strength can fluctuate throughout the course of a test session, so problems may arise during the session even if the initial systems check reveals no problem. To help ensure test takers can test through completion with no or limited disruptions, some recommendations in addition to the systems check include having the test taker clear the router before signing-in to test and limit others in the location (e.g., home) from being on the same network, in particular heavy bandwidth usage like gaming or streaming. A wired connection is preferred to a wireless one, and hotspots should generally be avoided as they are less likely to maintain a strong connection throughout the test due to the means by which Internet connect(s) are transmitted.

Firewalls can also create problems, especially if a test taker does not have administrator rights to the computer being used. The more secure the Internet connection with the test delivery provider, the more likely a test taker may encounter an issue with firewalls. Such problems are most common if a test taker is testing on a machine provided by an employer or other third party. Many systems checks will not fully test for the presence of firewalls and can create an issue when a test taker attempts to sign into the test.

Supporting Locations Without Reliable Access to Resources. There are several ways to address communities without reliable access to the resources that TBAs require. For example, while the world continues to advance toward digitization, reverting to paper-based testing (PBT) feels like a step backward. Nevertheless, there are instances where PBT is the safest and most reliable means to deliver an examination when the required resources for TBA are not accessible or strong. Relative to other solutions, PBT is often the most easily implemented and cost-effective solution for meeting the needs of some communities. In such cases, the psychometric comparability of these tests should be examined compared to their TBA counterparts when comparability is required by the testing purpose to ensure the exams are fair, valid, and legally defensible.

There are also options for deploying TBAs in locations without reliable resources or where access to the Internet is prohibited, such as in prisons or highly secure institutions. Even when the challenges of delivering a computer-based test can be overcome, the workload and costs may be substantial.

Disconnected delivery is one possible solution for these communities, where test content is downloaded and housed locally on a machine and then delivered to test takers. This delivery can be accomplished in either group settings or in a one-to-one setting. When used in a group setting, the investment in computer equipment for a sponsoring organization may be significant, as well as the management of the administration process, ensuring that all machines are working as expected and secured before and after the administration. Like PBT, this model carries a security risk in that the machines could be stolen or tampered with, creating a risk that test content could be compromised, which for some organizations could mean a substantial financial loss. Thus, in this model there should be rigor around protecting the machines and for a means to destroy the content(s) of the machines remotely if there is concern that machines have been compromised.

The most expensive model for supporting communities with scant technology resources is using a one-to-one proctor to test taker experience. In this model, a proctor would use a technology such as disconnected delivery to deliver a test one-on-one with a test taker. Both the test sponsor and vendor should be well informed of the laws and risks associated with such an event and plan the venue and time of the event accordingly to minimize risk to all parties involved. Many times, these one-to-one events are reserved for VIP test takers (such as CEOs of companies that must maintain licenses/certifications) or individuals with unique accommodation requirements.

Benefits to test takers and test sponsors for having optional delivery modalities available when Internet connectivity is not possible include the following:

- Fairness for test takers: Ideally, all test takers should be offered the same or similar opportunity to demonstrate competency and should not be treated differently due to lack of Internet service.
- Standardization of the test taker experience: All TBAs should be conducted with the same "look and feel" test features (e.g., randomization of test items, use of tools such as an online calculator, etc. should be available to all test takers).
- Scoring: Even in situations where there may be delays in uploading results files until an Internet connection can be established, having all test data in the same format speeds analysis and processing.

Candidate Preparation, Practice, and Orientation to the Technology

The guidelines for test taker preparation, practice, and orientation to the technology are designed to help ensure test takers' experiences are fair and test scores reflect the knowledge, skills, abilities, and other characteristics of test takers (i.e., safeguarding accurate measurement while minimizing construct-irrelevant variance). The goal of preparation, practice, and orientation to technology in cognitive ability testing is to prepare and motivate candidates so they can accurately demonstrate their performance level. The goal of preparation, practice, and orientation to technology in behavioral testing is to prepare and motivate test takers to respond with their most realistic and honest responses to the tasks. In all cases, test takers thereby should be given adequate opportunities to prepare, practice, and understand the testing technology prior to testing (Bishop & Davis-Becker, 2016; Zwick, 2006).

For test takers to accurately represent their standing on the constructs in a TBA environment, they must have a high level of comfort and familiarity with the technology (Llabre, Clements, Fitzhugh, & Lancelotts, 1987; Parshall, Spray, Kalohn, & Davey, 2002; Russell, Goldberg, & O'Connor, 2003). Early studies of computer-based assessments indicated test takers felt higher levels of anxiety when testing on a computer as compared to paper-and-pencil testing (Llabre et al., 1987; Ward, Hooper, & Hannafin, 1989). Although test takers today generally have higher levels of comfort with computer technology, some populations may not be familiar with the technology demands of the test, and even those who are, may experience high levels of anxiety when answering test questions using unique or unfamiliar item formats or computer interfaces (Bishop & Davis-Becker, 2016; Sireci & Zenisky, 2016).

Although educational and psychological measurement has broadened the spectrum of item tasks to improve measurement, test takers may lack familiarity and comfort with many technology-enhanced item formats. Consequently, it is incumbent on testing organizations who utilize innovative or technology-enhanced formats to provide test takers with explanatory descriptions and opportunities to practice. To minimize construct-irrelevant variance attributable to differences in test preparation and practice opportunities, testing organizations should provide all prospective test takers *free practice opportunities* along with *appropriate explanatory materials* of the testing interface and item tasks. These materials should also explain rules governing testing, such as time limitations and inappropriate responses. Organizations should provide for and encourage all prospective test takers to: (a) read and understand the rules of testing; (b) study explanations of item tasks; (c) use practice items and review feedback; and (d) take a practice test form in the actual test interface.

Appropriate Test Preparation. Questions arise regarding what constitutes appropriate test preparation and practice activities. These questions have become more pressing as accounts of large-scale cheating are reported in public schools (Chen, 2018), college admissions (Paris, 2020), and professional licensure testing (Lubin, 2013; Prometric, 2020). Appropriate test preparation and practice activities enable test takers to accurately demonstrate through testing their knowledge, skills, abilities, or other characteristics (Crocker, 2006; Lai & Waltman, 2008; Popham, 1991, 2003). As Popham (1991) offered, “No test preparation practice should increase the student’s test scores without simultaneously increasing student mastery of the content domain tests” (p. 13).

Applying this principle to the development of preparation and practice materials, testing organizations should build preparatory materials and practice exercises that enhance test takers’ understanding and comfort with the test technology, including the test interface, tools, and formats so that they can best demonstrate their standing on the construct(s).

Guidelines for Global Considerations in Technology-Based Assessments

Guidelines for Translation and Adaptation of Technology-Based Assessments

11.1 When needed, translation/adaptation should be planned as part of the test development process and assessment design.

Comments: Some testing involves assessing test takers who operate in different languages. When tests are to be administered in different languages, the development of the multiple language versions should be considered from the earliest stages of test development. It may be helpful to set up a multidisciplinary task force of test developers, test platform engineers, partners or experts from (a subset of) the target regions, and translation experts in planning the development process. Attention should also be paid to different measurement systems (e.g., metric, imperial) and currencies.

11.2 A testing organization should clearly define the constructs to be measured as well as the generalizability of these constructs in terms of comparability across different language versions of the assessment.

11.3 A process should be developed, implemented, and maintained to standardize and centralize documentation of each step of the translation/adaptation process, including all adaptation choices made for each item and for each locale.

Comments: Collect and retain item-per-item translation and adaptation notes based on the results of qualitative and empirical studies to make item and test revisions and to inform future test development and adaptation efforts. The translation and adaptation notes aim to give specific guidance to accurately translate stems, stimuli, or expressions to maximize psychometric equivalence to source versions. Such notes should specify when and how to adapt specific parts of the text. If possible, these notes should be available to translators and reviewers in the computer-assisted translation tool or translation environment they work in. If applicable, retrieve or create translation memories from previously existing translations of test items that are being recycled from a previous test administration. Organize a centralized repository for all translation-related resources. Consider using bilingual glossaries and style guides for each of the target locales. See additional information in 11.11 related to quality assurance.

11.4 Translation/adaptation and linguistic QA design should be established in consideration of time and budget constraints.

Comments: In many cases, a multistage team translation model will be helpful. All required team members should be identified, and provisions for hiring and training translators and reviewers should be provided. Note that back translation will only give you very limited information about the suitability of the target version for its data collection purpose.

11.5 A detailed documentation plan and list of communication channels should be included for each step of the translation, adaptation, and linguistic QA process.

Comments: This plan should include a translation/adaptation timeline and contingency plan, as well as evaluation of translation of language related to the testing platform and navigation.

11.6 Translation/adaptation processes should consider how to handle languages or cultural variations shared by different target regions.

Comments: Include an approach to harmonize the differences among the shared language across the versions within a target region.

11.7 Strategies to gather evidence to evaluate the translation/adaptation of test content should be incorporated throughout the test development process, and sufficient validity evidence should be provided for each language version of a test.

Comments: Such studies can include cognitive pre-testing of a subset of translated items, conducting focus groups to prepare protocols in the target languages, and conducting statistical analyses (such as differential item functioning and other measurement invariance studies). Studies should be planned to evaluate the comparability of scores across different versions. Performance and validity of items with features that pose challenges in translation/adaptation should also be evaluated.

11.8 Translated/adapted versions should be piloted when possible.

Comments: Statistical analyses on pilot data can inform possible revisions to the source version. Refrain from cosmetic or preferential edits after the pilot: they can affect psychometric properties, including validation, of items in unforeseeable ways.

11.9 Training should be provided for linguists, reviewers, subject matter experts, and other players involved in the translation process.

Comments: If possible, the test developers should be involved in the training. Provide technical support to linguists if they translate inside the platform and/or if they use the platform to preview translated materials. The translation vendor should provide support to linguists if they translate outside the platform; demand from the translation vendor that translation memories should be part of the deliverables. Translators should also be trained on test security and issues.

11.10 Suitability of the authoring platform should be investigated for the target languages/cultural variations envisaged and related technical challenges identified.

Comments: Consider exporting content from the platform and producing translations outside the platform (see translation data exchange standards). If possible, avoid integrating the translation process in the test authoring/delivery platform: it usually precludes harnessing the power of mature translation technology.

11.11 QA processes and translation quality checks should be performed.

Comments: These processes can include optimizing source content from a technical perspective, ensuring segmentation rules are correctly applied in the authoring platform, evaluating the extraction process in collaboration with a translation technologist, and evaluating the different translation workflows (including the export and import process). The QA processes should be collaborative across test developers and linguists. These checks can include both qualitative and quantitative evaluations. Documentation of these evaluations can support the validity of the translated/adapted assessment. Consider a separate process for a final layout check once the verified translations are imported back into the platform. Consider using an independent agency in the process.

11.12 The translation/adaptation workflow should be implemented and monitored by using a dashboard or similar tool to manage translation progress or require regular progress reports from the translation vendor.

Comments: It is recommended the test developer appoint an officer to examine requests for intentional deviations from the source version rather than outsourcing decision-making about (unforeseen) adaptations. Implement a selection of downstream linguistic QA processes. Gather standardized feedback and detailed reports on reviewer interventions and factor in one or several iterations to adjudicate controversial issues.

Guidelines for Testing Where Technology Resources Are Low

11.13 When Internet bandwidth may be limited, strategies should be used to optimize testing conditions, such as clearing out the router before sign-in, limiting others from being on the network, and using a wired connection instead of a wireless connection.

Comments: Hotspots should generally be avoided as they are less likely to maintain a strong connection throughout the test due to the means by which Internet connect(s) are transmitted.

11.14 Paper-based tests may be appropriate substitutes for TBAs when there are inadequate technology resources for delivering the test.

Comments: Evaluating technology resources is a necessary step not only for the initial implementation of TBAs but also as part of the preparation for routine operational test administration. In cases of test interruption due to inadequate technology support, consider using PPTs as a replacement. In such situations, research will be needed to ensure the PPTs sufficiently fulfill the same purposes as the TBA. If scores will be compared across paper-based and digitally delivered assessments, validity evidence of score comparability will be needed.

11.15 When digital test results are collected from remote servers, the data transfer should occur immediately after the completion of a test or after the completion of each item for online (e.g., Internet-administered) tests. The data should be protected by strong access procedures while residing on a remote server and strong encryption during transmission.

Comments: When administrating TBAs in environments when results cannot be immediately transmitted, processes should be in place to ensure the results are stored safely and in a manner that allows for later retrieval.

11.16 When test content is distributed, whether in booklets or digital file form, it should be protected at every step of the distribution process and stored securely at testing locations.

Comments: Regardless of the method used in providing TBAs in locations with limited or no technology resources (i.e., no Internet connectivity), test content needs to be secured and *protected at every stage of the process. During test administration and after authentication, tests are at greatest risk. For example, during this time, displayed items can be stolen, and other forms of cheating may occur. In addition to prior agreed-upon efforts during the planning and design stages, additional effort may be needed to ensure, to the extent possible, test content cannot be stolen, and the probability of cheating is minimized. Test administration in locations where resources are limited or not available may inherently be riskier as security provisions used when resources are available cannot be employed to protect test content.*

Guidelines for Test Taker Preparation, Practice, and Orientation to the Technology

11.17 Information about the purpose of the test, test registration, test content, item formats, and item scoring should be available to all test takers well in advance of testing in an easy-to-access medium.

Comments: All prospective test takers should have access to general information regarding the test, including the description of the purpose of the test, description of test score use, privacy protections (see Chapter 9), and test administration information. Such information should include when and where testing occurs, registration requirements and restrictions, personal identification requirements, instructions for taking the test (including how to respond to test items), materials and aids that can and/or should be brought to testing, materials, aids, and behaviors prohibited during testing, retest policies, and clear explanation of the consequences stemming from violation of test administration rules. Score cancellation policies in place should also be clearly communicated. A "test taker agreement form," signed by the test taker to confirm receipt of this information, may be helpful in the event of any disputes. Proctors and other actors involved in testing should also have access to these materials.

11.18 All prospective test takers should have access to general content information (i.e., what is covered on the test), except in cases when such information affects measurement of the intended constructs.

Comments: Information should be provided to describe the content domain tested (behaviors, knowledge, skills, abilities, attitudes, or other characteristics), content that is not measured on the test, the item formats utilized, and how test items are scored (scoring rules and rubrics for each item format). Examples of item formats, including providing samples of exemplary responses, should be provided.

11.19 Information regarding the test interface and hardware/software requirements for testing should be available to all test takers in an easy-to-access medium.

Comments: All prospective test takers should have access to specific information regarding the test interface, including clear descriptions of all features provided by the test interface, screen shots of the test interface with its various features identified and described, examples of test takers using *the interface features to successfully navigate through the test, and information and tips for successfully navigating and using the interface.*

11.20 All prospective test takers should have access to a list of minimal hardware and workspace requirements for testing.

Comments: The testing organization should require test takers to check the specific hardware feature (e.g., camera, microphone) during registration. If software programs are required for testing, prospective test takers should be made aware of the requirements prior to registering.

11.21 Preparation and practice opportunities should be conveyed to all test takers in an easy-to-access medium.

Comments: If preparation or study materials have been developed, information on how test takers can obtain/purchase these materials should be conveyed in an easily accessible manner. Test takers should be encouraged within either the general test information or in the registration materials to study and review the practice items. Ideally, practice items should provide feedback to test takers.

11.22 Prospective test takers should have the opportunity to take a full-length practice test.

Comments: The full-length practice test should be presented in the test interface and should be built to the same specifications as the actual test. Test takers should have the opportunity to complete some or all of the practice test. After completing the practice test, prospective test takers should receive feedback regarding their performance.

11.23 Information about score interpretations and reports should be provided to all test takers prior to testing.

Comments: Prospective test takers should have access to information regarding scores generated through the test. Sample score reports should be available in an easy-to-access medium, including descriptions of the information conveyed through the score report. The meaning and interpretation of all scores and subscores should be explained, and clear information regarding who receives the score reports should be provided.

PART IV. EMERGING APPLICATIONS IN TECHNOLOGY-BASED ASSESSMENT

Emerging applications of technology and data science in assessments are rapidly evolving and have not yet reached state of the art in practice where appropriate guidance can be offered at this time. Included in this section are brief discussions of artificial intelligence (AI) and examples of emerging technology-based assessment applications used to assess and make decisions about people, such as mining "big data" sets and social media data, facial recognition and analysis, and some types of automated generation of test items.

These innovative applications have begun to appear in assessments (Oswald, 2020; Weiner & Foster, 2018; Zickar, 2018). However, accepted best practices are not yet settled, so it is too early to form consensus guidelines for the design and application of AI-based assessments, although related data privacy considerations are discussed in Chapter 9. Instead, this section provides discussion of some of the emerging issues and provide references to other relevant documents. Considering the increasing pace at which these innovations are being researched and applied, we anticipate that guidelines will be developed for these applications in the not-too-distant future. The Association of Test Publishers (ATP) released a white paper containing additional discussion of issues pertaining to AI in assessment (ATP, 2021) and has published a set of principles for testing organizations in the development and use of AI systems (ATP, 2022).

Artificial Intelligence

The availability of diverse, cross-referenced, and increasingly large data sets (big data) and the continued scaling and availability of cost-effective computation cloud computing has made possible have combined to increase the application of AI in assessment. This application has enabled advanced statistical analysis practices ("advanced AI"), including machine learning (ML) systems, to identify novel applications and predictions from once intensely manual, unanalyzable, or impractical assessment scenarios. Examples include the use of adaptive testing and learning systems (see Chapter 2) and the use of ML and natural language processing (NLP) in modeling and scoring constructed-response assessments (see Chapter 4).

Significantly, however, the definition of AI in some legal and regulatory contexts has been limited to exclude traditional software merely used to automate human actions rather than substitute for human decision-making.[10] Examples of this non-AI approach include automated test scoring applying the same

[10] A "compromise text" to the draft AI Regulation in the Artificial Intelligence Act released by the European Council (November 2021) clarifies that traditional software that merely automates a manual task is not considered AI, in contrast to a system that requires data learning, reasoning, or modeling to reach outcomes. Thus, some testing

rubric used by human scorers and automated item/test construction for test delivery. While such automated decision-making requires privacy attention under the EU's General Data Protection Regulation, the ATP has taken the position that automation software should not be classified as AI.

An AI system is "a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy" (Organization for Economic Co-operation and Development - OECD, 2019).[11] During over 60 years of research and experimentation, AI has shifted from an interesting field of study to a driver of global economic growth and a strategic priority for almost every organization and industry. The assessment world is beginning to incorporate advanced AI, and it is expected to become increasingly important in this arena, offering advantages such as scalability and efficiency in assessing higher volumes of test takers, use of data-rich measurement models, and greater fidelity through technology-based simulation of task performance.

AI may be characterized by five core principles (OECD, 2019). AI (1) generates a predicted output from input, using historical "experience" data; (2) provides a measure of the confidence the system has in its prediction and enables the determination of required for action; (3) requires significant computing power to arrive at a prediction in a timely fashion; (4) requires a large corpus of material or an environment for repeatable experiments to build experience, against which prediction can be made; and (5) has a stronger "understanding" of probability than most people have.[12] TBAs may leverage these AI principles by using historical test taker data to develop and "train" models to replicate and exceed the capacity of humans; for example, in generating test items, analyzing test taker responses to detect potential bias, or scoring complex data, such as essays or trace data from a game-based assessment.

AI contains many fields of research including ML, expert systems, NLP, and other domains. The field of AI has evolved in three general waves: (1) Symbolic AI, rules, logic, and data that are broken down into decision trees; (2) Boolean criteria and outcomes, probabilistic expert systems, and weighted trees; and (3) most recently, learning systems, including ML, deep learning and various offshoots using complex data algorithms. The latter types of AI systems are the ones drawing serious attention and criticism because of concerns over the potential existence of bias and discrimination in the AI system or the use of PI to make decisions. In the field of assessment, there are many examples of early AI techniques being

software used today (e.g., scoring, item generation, test monitoring) should not be considered or treated as AI for regulatory purposes.

[11] There are many current proposed definitions of AI, including some for legal purposes. However, until a formal definition is agreed upon, the testing industry believes the most appropriate definition is where the AI system engages in "learning, reasoning, or data modeling."

[12] Of critical importance to this discussion, "prediction" by an AI system, and the need to evaluate whether bias exists in an AI system, are not the same concepts as used in psychometrics (e.g., validity, reliability, and fairness as those terms are used in the *Standards for Educational and Psychological Testing; AERA/APA/NCME 2014)*. Thus, it is necessary to consistently distinguish between the role of psychometrics and AI in testing. The ATP comments on the European Commission's draft AI Regulation emphasize the need for this distinction (ATP, August 2021).

applied, now recognized not to constitute advanced AI. Many of these techniques date back to the early symbolic period, for example, with automated scoring.

For assessment providers and organizations, opportunities to explore and apply AI, big data, and social media inputs abound (Oswald, 2020; Zickar, 2018). Some forms of adaptive scoring, dynamic baselining, real-time behavior modification, and fraud detection are examples of delivery-oriented opportunities, while encryption, complex item authoring, normative calculations, and variance detection, as well as novel new test types, exemplify design and development opportunities open to all types of assessment providers. It is important that users of AI be responsible and accountable for the decisions to assure that an AI system promotes fairness and trust by eliminating or minimizing bias/discrimination.

Big Data and Social Media

New sources of unstructured data, including PI, continue to increase in volume, variety, and velocity, driving the incorporation of historical and real-time data into analysis, insight, and action in potential assessment applications (Oswald, 2020). Big data applications are evolving. What was considered "big" in the early years of the last several decades, which drove advances in data storage and accessibility (e.g., Hadoop, DFS, NoSQL), has given way to even larger scale analysis of events and experiments that continue to push the boundaries of storage and computing, including the use of neural networks and quantum computing.

Beyond being a significant source of big data, social media has two potential contributions to assessment: visibility into the social graph of an individual and insights into behaviors, perspectives, and opinions. Social media provides dynamic data for potential use in assessment, complete with feedback loops and opportunities for individuals to adjust and modify their responses in real time. The exploration and use of social media data for assessment is in its nascent stages at this time (Zickar, 2018). However, it is important for testing organizations to consider how to balance these uses with applicable legal privacy requirements (see Chapter 9).

Facial Recognition and Analysis

The use of facial recognition and analysis of digitized facial data is widespread in security applications such as personal identification, video surveillance, and secure access to systems and devices (Klosowski, 2020). Digitized facial data have also been increasingly used in actual assessments, such as performance in video interviews and exercises (Oswald, 2020). Facial analytics (FA) is the application of AI being leveraged to replicate human vision, which incorporates recognition and machine vision capability.[13] FA may be used in video interviewing with recorded asynchronous interview scoring, while other

[13] Notably, FA is only one example of the use of biometric data for purposes of identification or in AI to enable profiling of individuals and/or individual behaviors. Biometric data also may involve fingerprinting, iris scanning, vein/palm scanning, voice recognition, handwriting analysis, and other ways for AI to profile a person's physical features. Significantly, here again, the main legal issue is the privacy implications of using personal information to make AI-generated decisions about the individual.

applications purport to derive information about communication skills, as well as sentiment, emotion, and personality attributes. FA may also be used in test security to authenticate the identity of a test taker and to monitor test taker behavior during a test (e.g., use of gaze detection to identify the amount of time a test taker is looking away from the computer monitor or has left the testing session).[14]

FA uses ML to detect a person's face. The machine is trained to detect facial landmarks, such as eyes, brows, and lips. By comparing these landmarks between sources, it can verify the identity of a test taker. These landmarks and features are combined to create a simplified model of the person's face that can be used in training the machine for the final step – feature classification. Here, ML algorithms are trained to classify feature groups based on images submitted with a known emotion or facial expression. For example, thousands of pictures of people showing positive sentiment (smiling) are submitted along with the defined set of features to create a deep learning algorithm that is able to classify those feature combinations.

The use of FA in assessment requires careful consideration of privacy laws and regulations (see Chapter 9), as some jurisdictions heavily restrict the use of biometric data. Fairness and bias issues have been observed with AI algorithms (see Chapter 10), which are also a concern as certain applications of FA have been shown to be unequally effective with all skin colors/tones, and some classes of test takers may exhibit facial expressions that do not align with normative interpretations (e.g., in the case of neurodivergence or visual impairment). Thus, the use of FA in assessment requires cautious design, monitoring, and evaluation in implementation (Tippins, Oswald & McPhail, 2021).

Automated Item Generation

Technology and AI-based methods are also being used to create content, especially for high-stakes high-volume testing programs that require large numbers of test items to mitigate security risks and avoid overexposure of test content. Automated item generation (AIG) is a method that helps to address these concerns. Traditional non-AI AIG approaches generate items from a model or template by substituting words and/or numbers that are intended to change the question, while measuring the same knowledge, skill, ability, or characteristic. However, some recent models attempt to use AI to generate items directly from a corpus of information. AIG is an evolving concept that has been the subject of research among psychometricians (Gierl & Lai, 2013). Empirical research and cost-benefit analyses continue to be explored in the psychometric community. A fundamental challenge in AIG is determining which modifications can be made without affecting the psychometric properties of a test item (Yaneva et al., 2020), or to be able to model and explain psychometric properties (Attali, 2018; Cole et al., 2020).

[14] However, if the use of facial recognition technology does not authenticate a person's identity, but is limited to verifying that a test taker is the person who registered to take a test (i.e., matching a person's face with a previous image provided by the test taker, or one shown on the computer camera at the start of a testing event), there is reasonable argument that this software should not be deemed to constitute AI since the technology provides only a one-to-one match against a known subject, and does not involve the use of algorithmic AI software to determine who the person is from among a multitude of possible individuals.

Regardless of the AIG model used, it is important to incorporate quality assurance processes to ensure that the AIG algorithms are generating content as intended.

Regulatory Considerations

The capabilities of AI systems have grown through access to greater and more diverse data sets, development of stronger heuristics, and the application of increased computational capacity to any given use case. At the same time, societal, regulatory, and governmental interests and concerns over individual privacy and data security have grown. As a result, Responsible AI and Trustful AI initiatives have been introduced, and a myriad of data use and privacy policies and regulations have been enacted or are in progress (see Chapter 9). AI regulations are evolving, particularly in Europe, Canada, and the United States. At the time of this writing, issues have been brought to the forefront concerning transparency, privacy, bias in data, predictions, and decisions based on automated systems and AI. (Hind et al., 2018; Bender & Friedman, 2019; Gebru et al., 2020). Some preliminary regulatory proposals would declare that virtually every assessment used in employment and education is a high-risk activity, which would require both the developer and the user of an AI system to meet burdensome regulations to prove the AI system is not biased and does not discriminate (EC Proposed AI Regulation, US National Conference of State Legislatures, 2021).

Akin to the controversy over automation software being eliminated from the legal/regulatory definition of AI, a similar question has arisen over the use of biometric data. The EU Commission promulgating the AI Regulation has clarified that the use of biometric measures for verification/authentication purposes (i.e., to confirm that a specific person is who s/he claims to be or having access to services, building, or device), would *not be prohibited* "because such systems are likely to have a minor impact on fundamental rights of natural persons compared to biometric identification systems which may be used for the processing of the biometric data of a large number of persons." As a result, the EC has amended the proposed definition of "biometric information system" Art 3(36) of the AIA.[15]

The ATP monitors and provides regular updates to the industry on laws and regulations, and it is anticipated that ATP will provide separate documents pertaining to AI regulations and principles. ATP submitted comments on AI regulation to the European Commission (ATP, 2021). An initial document setting forth AI principles for testing organizations has been published (see ATP, 2022).

Conclusion

Technology has become an essential part of assessment throughout the testing lifecycle and holds promise for its continued evolution to achieve greater capabilities. This is true in education, employment, credentialing, and clinical assessment. As technology advances have transformed testing, fundamental concerns remain the same with respect to ensuring assessments are valid, reliable, fair and

[15] Regular updates to this and other AI regulations may be obtained from the ATP website, including ATP published comments.

unbiased, accessible, and secure, without introducing irrelevant variance in scores or unintended consequences. To this end, these *Guidelines for Technology-Based Assessment* provide information about key factors, issues, and best practices that should be considered when designing, delivering, and scoring tests using digital platforms, with the aim of ensuring fair and valid assessment.

GLOSSARY

Accessibility Support: A modification to the standard manner in which an item is presented, interaction with item content occurs, and/or a response is produced that is designed to improve an item's ability to collect evidence regarding the construct the item is intended to assess.

Accessible Portable Item Protocol (APIP): A standard format for tagging alternate and supplemental item content designed to support specific accessibility needs.

Accessible Rich Internet Application: Technical specifications that allow content developers to associate verbal representations of content elements and specify the order through which users of assistive communication devices navigate content.

Adaptation*:* Intentional deviations from a source language version of an assessment, made to conform to local usage or context. This term can apply to same-language versions of a test to be administered in different cultures or regions; it can also apply to deviations from the source version when translation would potentially put the respondents from the target group at an advantage or a disadvantage.

Adaptive Instructional System: An artificially intelligent, computer-based system that guides learning experiences by tailoring instruction and recommendations based on the goals, needs, preferences, and interests of each individual learner or team in the context of domain learning objectives (Sottilare, 2020).

Algorithm: A software program to handle a variety of computational tasks. Simple algorithms, more commonly referred to as "automation software", merely automate traditionally manual tasks by following established rules or routines and are not usually considered to be Artificial Intelligence (AI) (e.g., automated scoring, some forms of automated item generation), whereas more complex algorithms usually involve the use of AI.

Alternate Representation: A version of content that presents the same information in a different form, such as a braille representation of text.

Anonymization: Removal of all personally identifiable information so that information cannot be associated with a specific individual; anonymized data should not be able to be re-identified, although in some instances, anonymized data may be identified if a statistically invalid sample of individuals is considered. Anonymized data is usually aggregated when used for research purposes (e.g., norming, test improvement).

Artificial Intelligence (AI) or AI systems: Software and/or hardware systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge or processing the information derived from this data, and deciding the best action(s) to take to achieve the given goal. AI systems can be used in technology-based assessments to assist humans in making test administration and scoring decisions or to make automated decisions in place of humans.

.

Assessment Data: Data collected from learner or examinee interactions and aggregation of data collected from learner or examine interactions that contribute to the evidence base required to make an inference of attainment of knowledge, skills, and attributes of interest.

Assessment: Evidence aggregation across interactions to make an inference of attainment of knowledge, skills, and attributes of interest.

Assistive Technology/Assistive Communication Device: Software and hardware that provides access for the special needs of some users that directly access application content and allows for user input. Examples include screen readers, refreshable braille displays, sip-and-puff devices, and switch buttons.

Automated Item Generation (AIG)*:* Templating, cloning, applying automated software, or in other settings, applying AI cognitive modeling processes, to generate a set of items from a single input or scenario created by an item writer. Some forms of AIG do not incorporate AI, while emerging methods have begun to utilize AI.

Automated Scoring*:* A technological method that involves response matching or text/natural language processing to review and evaluate text responses in a reproducible way that matches defined scoring rubrics and is in agreement with human raters. Many forms of automated scoring merely use non-AI software to perform a manual task; however, in some settings, AI-based scoring systems are used and are continuing to evolve.

Behavioral Test: a test designed to measure an individual's tendencies to respond in particular ways to specific circumstances. In behavioral testing, items do not have right or wrong answers. An example of a behavioral test would be one measuring the Big Five Personality Traits.

Big Data: A volume, variety, veracity, and velocity of data that can be used by an AI system to train on, learn from, or reason against.

Cognitive Ability Test: a test designed to measure an individual's abilities to perform various mental activities involving processing, acquisition, retention, conceptualization, and organization of sensory, perceptual, verbal, spatial, and psychomotor information (AERA, APA, & NCME, 2014).

Comparability/Score Comparability: The degree to which similar inferences can be made across different variations of an assessment procedure, such as a parallel form, accommodated test administrations, or test delivering platform. Test linking can be used to facilitate score comparability, with the degree of comparability resulting from a linking procedure varying along a continuum that depends on the type of linking conducted.

Computer-Administered Test: A test administered by computer; test takers respond using the keyboard, mouse, and other technological devices (AERA, APA, & NCME, 2014).

Computer-Adaptive Test (CAT): A form of automated testing where the test taker receives successive items, or sets of items, which are selected in relation to the test taker's responses to previous items, in consideration of psychometric and content information.

Computer-Assisted Translation Tool*:* A broad term that can encompass any computer software used by human translators during the translation process to improve their working conditions and increase translation quality (Bowker, 2002). The term CAT tool refers to a computer environment that: (1) supports the translation of different file formats; and (2) allows the user to use and create language assets (e.g., terminology databases, translation memories).

Construct-Irrelevant Variance (CIV): Variance in candidate scores that is attributable to extraneous factors that distort the meaning of the scores and decrease the validity of the proposed interpretations (AERA, APA, & NCME, 2014).

Content Domain: The set of behaviors, knowledge, skills, abilities, attitudes, or other characteristics to be measured by a test (AERA, APA, & NCME, 2014).

Contextual Metadata: Metadata that allows for systems to interpret what data means on the source side and how to interpret the results on the output side.

Data Breach: A confirmed breach of security resulting in the accidental or unlawful destruction, loss, alteration, unauthorized disclosure of, or access to, personal data. Note: What constitutes a data breach may vary by law depending on the jurisdiction.

Data Controller (or **Controller**): An organization that, alone or jointly with others, determines the purposes and means of the processing of personal data. Typical examples of data controllers might be bodies that set certification exams, employers who test personnel or job candidates, or educational institutions testing students to make admission decisions or to score tests on course work.

Data Fitness for Purpose: How well data meets intended operational and decision-making goals, including freedom from defects and possession of desired features (Juran & Godfrey, 1999).

Data Forensics: In the field of assessment, data forensics pertains to the application of statistical methods to detect anomalies in test taker response patterns and test data to identify potentially serious test irregularities (e.g., cheating, proxy testing, content theft, and infringement of intellectual property rights).

Data Governance: The exercise of authority and control (planning, monitoring, and enforcement) over the management of data assets (DAMA International, 2017). Policies and best practices that ensure data is managed properly.

Data Lake: A system that acquires data from multiple sources in an enterprise in its original form and may also have internal, modeled forms of this same data for various purposes. The data may be any type of information, ranging from structured to completely unstructured data. A Data Lake is expected to be able to derive relevant meanings and insights from sored information using various analysis and machine learning algorithms. (John and Misra, 2017).

Data Lineage: A description of data's origin, movement, transformations, characteristics, and quality that allows for an understanding of where data originated, how it is transformed, and how it moves into, across, and outside an organization.

Data Processor (or **Processor**): An organization that processes personal data on behalf of a controller. Typical examples of processors might be services companies that provide assessment or analytic services, test publishers that provide tests for an employer to use, or proctoring companies.

Data Stream: A data stream is a continuously fed exchange channel that is part of a streaming data system: "a non-hard real-time system that makes its data available at the moment a client application needs it. It's neither soft nor near--it is streaming." (Psaltis, 2017).

Delivery Modality: The means by which an assessment will be delivered. Delivery modalities may include standard computer monitors and handheld devices, such as a cell phone.

Digital Badge: Electronic symbols used as credentials (or micro-credentials) to document achievement or skills mastered such as course completion, professional development participation, or training completion (Parker, 2015).

Disconnected test session: When test content is cached a number of items up front, e.g., a section at a time to mitigate temporary disconnects.

Distributional Equivalence: Similar score distributions across modes, devices, and technologies.

Downstream Linguistic Quality Assurance*:* Components of the linguistic QA design that take place after the actual translation, e.g., translation verification, final layout check, etc.

Drag-and-Drop: A technology-enhanced item format in which graphic tokens are dragged and dropped onto targets to respond to a query.

Elastic Computing Cluster: The dynamic provisioning of computing resources (e.g., virtual servers) using a system that allocates and reclaims CPUs and RAM in immediate response to the fluctuating processing requirements of hosted IT resources (Erl et al., 2013). This allows a cluster to automatically scale up or scale down based on load/computing needs.

Equating: A process for relating scores on alternative forms of tests onto a common scale so they have essentially the same meaning and facilitate comparable test score interpretations. The equated scores are typically reported on a common score scale.

Evidence Aggregation: The summarization of discrete pieces of information related to the knowledge, skills, and attributes of interest.

Extraction: The process of deciding which parts of the document are translatable and which parts are not. Ideally, all elements that should not be translated should be protected or hidden.

Game: "A system in which players engage in an artificial conflict, defined by rules, that results in a quantifiable outcome" (Salen & Zimmerman, 2004, p. 80).

Game Loop: A repeatable sequence of actions that players engage in to advance in a game.

Game Mechanic: The actions players take in the game world, how they do them, and the game response that results in progression through the game.

Gamification: The application of game elements to non-game situations. Notably, this does not have to mean just points and leaderboards but can also include a variety of game elements, including narrative, quests, social features like guilds, levels, and boss battles.

Historically Marginalized: A term referring to groups of people who have been consistently, repeatedly, and deliberately excluded by the wider society. Although the term is often used in relation to economic or political opportunities, or lack thereof, this marginalization occurs across multiple sectors (e.g., education, health [von Braun & Gatzweiler, 2014]).

Hotspot: A technology-enhanced item format in which the test taker responds by clicking directly on an image.

Incomplete Testing Session: An item response pattern with some missing item scores for a test taker due to technological disruptions.

Informational and Interpretive Materials to Support Results Reporting: Materials that typically encompass text-based resources to support the use of results accessible in a reporting portal of some kind such as (but not limited to) plain-language interpretive guides, technical documentation for scores and testing programs, and perhaps a frequently asked questions document.

Interaction: Observable piece of data that provides information about a user's interaction with a system.

Interactive Reporting: A high degree of user choice in determining what results and/or analyses are called up to be displayed on a web page or included in a report on the fly. Such efforts typically involve group-level reporting (at a scale and grain size at the discretion of the user).

Interoperability: The ability of systems or software to exchange and make use of information.

InterpretML: An open-source library of machine learning explainers and interpretability techniques for explaining an AI model in an interpretable manner.

Item Bank: An electronic data file containing test questions, item content, attributes, and metadata. See item pool.

Item Format: The way in which the task or question for test takers is presented and the way in which the test taker provides a response.

Item Pool: An electronic database of test item content, associated attributes (e.g., scoring key, content classification, cognitive level, enemy items) and metadata (e.g., item statistics, historical use), from which test forms may be drawn manually or automatically (in the case of linear-on-the fly testing, LOFT) or items many be selected individually for test delivery (in the case of CAT).

Item Response Theory (IRT): A theory of testing based on a mathematical model of the relationship between performance on a test item, the test item's characteristics, and the test takers' levels of performance on the construct being measured. Different statistical models may be used to represent item and test taker characteristics.

Linking Scores: A process used to relate scores across different tests.

Locale*:* Language-country combination (e.g., Spanish for Mexico or English for Singapore).

Linear-on-the-fly Testing (LOFT): An automated test assembly method that is used to assemble a unique equivalent form of a test to each test taker, drawing from a pool of items representing content domains and calibrated with respect to psychometric properties and other item attributes that are used to guide assembly.

Machine Learning (ML): A form of artificial intelligence that makes predictions from data. ML entails the use and development of computer systems that are able to learn and adapt without following explicit instructions by using algorithms and statistical models to analyze and draw inferences from patterns in data.

Masking: A technique often employed for individuals with information processing needs that reduces the amount of content presented on a screen or paper-based page by temporarily blocking select elements of content in order to support an increased focus on content that is visible.

Metadata: Data about data. More formally, characterization of the structure, content, and quality of data, including source and lineage and the definition and intended uses of entities and data elements (DAMA International, 2017).

Minoritized: A term different from the noun minority, referring to populations/communities of people who have less power or representation compared to other groups as a result of social constructs (Benitez, 2010). The verb minoritized more accurately describes the oppressive context in which these populations of people must exist and recognizes that systemic inequalities--such as racism, ableism, sexism, nationalism, etc.--have placed them into "minority" status through no control of their own.

Multimedia: Any visual enhancement to an item, including static images, video clips, audio clips, live video responses, and so on.

Multistage Test: A form of adaptive testing, similar to CAT, wherein sets of items are delivered to the test taker on the basis of the person's preceding responses to a set of items.

Natural Language Processing: A branch of artificial intelligence, linguistics and computer science in which computer software is used to analyze and "understand" written and spoken human language.

Offline: When the full test content is downloaded up front to allow for the test to be completed without an Internet connection.

Online: When test content is loaded in real time, one item at a time, (technically) limiting exposure of content.

Parallel Test Forms: Alternate forms of a test that are exactly equivalent, i.e., measure the same construct(s) and have the same means and standard deviations.

Personal Data (or Personal Information, or PI): Any information relating to an identified or identifiable natural person (also sometimes referred to as a data subject or test taker). Among the examples found in various national privacy laws and regulations, PI would include the person's name, address, IP address, national identification number/social security number, payment (card) information, and even some types of pseudonymized identifiers that are capable of being re-identified.

Personal Needs Profile (PNP): An extension of the APIP standard that allows users to specify the accessibility supports required by a given test taker. PNPs are used in APIP using Access for All (AfA) v2.0 and in QTI 3 using AfA v3.0.

Processing: Any operation performed on personal data, including but not limited to collection, recording, organization, structuring, storage, retrieval, using, transmitting, disseminating, or making the data available, as well as restricting, erasing, or destroying the data.

Pseudonymization: The processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organizational measures to ensure that the personal data cannot be reattributed to an identified or identifiable natural person (i.e., re-identified).

QTI: The Question and Test Interoperability specification, which defines a standard format for tagging item content and specifying the manner in which responses are collected and processed.

QTI 3: A specification that integrates APIP into the QTI specification. It includes updated accessibility supports (HTML 5, WAI-ARIA, Access for All 3.0), web-component friendly markup, and integration with Portable Custom Interactions (interactions developed for a specific scenario) and the standard on CAT.

Refreshable Braille Display: An electronic device connected to a computer that contains multiple cells, each with six pins that elevate or are depressed to produce braille characters. The characters displayed on the device are relayed from a test delivery system to present text content in a braille form.

Recommender System: A type of AI machine learning system designed to leverage content and person-specific metadata to predict or provide personalized recommendations. In a consumer-oriented context, recommendations can be products or services, often relevant to online search-related behaviors. In a technology-based assessment context, the recommendations can be of items and learning content. The purpose is to leverage far more metadata than traditional computer-adaptive assessments, leading to greater depth and diversity of reported information, ultimately extending performance results to instruction and learning.

Score Equivalence or Interchangeability: Scores regarded as equivalent in terms of construct and precision and that have the same meaning for the population.

Security Incident Response: Actions taken by a testing organization in response to a security incident using a pre-set written response plan, to investigate what happened, determine if a data breach occurred, and any remediation steps that should be taken.

Segmentation*:* The process of splitting a text into small, manageable parts--usually sentences.

Simulation: A system or subsystem that emulates or offers in a controlled fashion a recreation of a reality.

Social Media: A very large source of data and a real-time source of data that an AI algorithm can be trained on, can learn from or can reason over.

Source Version*:* The version of an assessment that serves as the starting point for the translation or adaptation. The source language is the language in which the source version has been developed.

Speededness: The situation in which the time limits on a standardized test do not allow substantial numbers of test takers to fully consider all test items.

Static Reporting: Reporting that includes results in tables, charts, and/or text formats generated by the test developer or other reporting agency that website users cannot manipulate. These may be available, for example, as downloadable PDFs that package pre-specified information in easy-to-print formats for user review.

Sub-Processor: A processor that works on behalf of a processor rather than directly for the controller. A typical example of a sub-processor might be a data center or data hosting company providing services to a processor. Sub-processors can have their own sub-processors and so on.

Target Version*:* A translated or adapted version of an assessment produced to measure the same construct or domain in a given target population. A target language is the language into which a target version has been translated or adapted. In case of adaptation, the source and target languages can be the same.

Technological Disruption: An event that disrupts the test administration experiences of test takers caused by the malfunctioning of hardware or software through which data are captured or transmitted, including hardware and software with which students interact directly, hardware and software owned and operated by the assessment provider, and hardware and software owned and operated by third parties that transport data between assessment provider and student. Examples of some common types of technology-related disruptions include delayed login, slowing down of the online system in the middle of the test, not receiving a second-stage test upon submission of answers to the first-stage test during a two-stage testing, being unexpectedly logged out, and losing some or all answers. See additional information below for Testing Disruption.

Technology-Enhanced Item (TEI): A test item that incorporates media or additional functionality that is only available through electronic means. TEIs are computer-delivered and require test takers to interact with the content in ways beyond selecting a correct response and provide a more authentic and engaging experience than traditional multiple-choice items.

Test Developers: Those contributing to test content or designing and maintaining the test platform and delivery system, who are responsible for test creation and delivery.

Test Disruption/Interruption: Any incident that occurs during the test administration that, from the point of view of the examinee, results in significant time delays, inability to enter or complete an assessment, or loss of examinee response data for one or more test items or tasks. This includes an event that disrupts a test taker's experience, caused by the computers, online systems, or other technological devices through which the test is delivered (Martineau, Domaleski, Egan, Patelis & Dadey, 2015).

Test Irregularity: Any event, including a technological disruption, that interferes with the standardized administration of a test, including its delivery and any proctoring services that are used. Examples of irregularities include disruptions affecting the entire group of test takers (e.g., a power outage, distraction noises such as fire alarms/sirens) or may focus essentially on a single test taker or several specific test takers (e.g., use of prohibited aids, such as a calculator or scratch paper, or even evidence of electronic aids or a camera, or disruptive behavior, such talking, passing notes, or interactions among several test takers or with a proctor).

Test Orientation: Test preparation activities that specifically include information about the structure of the test, test interface, time limits, item formats; preparation and practice with test-allowed tools (e.g., calculators, rulers, notes) (Allalouf & Ben-Shakhar, 1998).

Test Preparation: Activities specifically undertaken to (a) review content likely to be covered on the test and (b) practice skills necessary to demonstrate knowledge in the anticipated format of the test (Bishop & Davis-Becker, 2016; Crocker, 2006).

Test Taker: The person taking an assessment. Also known as an "examinee," or in the context of credentialing/certification exams, the "candidate."

Test User: An individual or entity who employs an assessment for a particular purpose. May also include a test administrator or proctor.

Validity: The degree to which the use of a test for a particular purpose is supported by theory and empirical evidence.

Web-Based, Internet, Internet-Based, Or Online Testing: Testing in which the Internet is the dominant technology for test administration. Through a continuous Internet connection, items are streamed as needed to a digital device used by the test taker. Each student's response is also returned immediately through the Internet to a server (Foster, p. 236).

Whiteness: A "quality derived from and against those 'Others' whom it sets apart as political, anti-individual and always raced (Barnett, 2000). Whiteness (like all notions of race) is fundamentally a relational concept rather than something residing *in* an individual or group" (p. 10). Whiteness--by virtue of being white alone--holds social, legal, economic, and political rights unavailable to others. Whiteness can maintain its power by declaring itself normal and good; and, consequently, solely worthy of its benefits. It is important to note that whiteness, as used here, is not about individual white people but about a political and economic social order (Sensoy & DiAngelo, 2012).

Wireframe: A layout of a screen used in the design and development of new games that demonstrates the elements that will be built into the game**.**

REFERENCES

Abedi, J. & Ewers, N. (2013). *Accommodations for English learners and students with disabilities: A research-based decision algorithm*. Smarter Balanced Assessment Consortium.

Advanced Distributed Learning Initiative (2019). *Total Learning Architecture 2019 Report*. (Retrieved Nov 12, 2020, from Https://adlnet.gov/assets/uploads/2019%20Total%20Learning%20Architecture%20Report.pdf).

Allalouf, A., & Ben-Shakhar, G. (1998). The effect of coaching on scholastic aptitude tests. *Journal of Educational Measurement, 35*, 31-47.

Allalouf, A., Gutentag, T. & Baumer, M. (2017). Quality control for scoring continuously administered tests. *Educational Measurement: Issues and Practice.* 58-68.

Allalouf, A., Hambleton, R. K. & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement, 36(3)*, 185-198.

Allalouf, A. & Hanani, P. (2010). Test translation and adaptation. In E. Baker, B. McGaw & P. Peterson (Eds.), *International encyclopedia of education (3rd ed),* 166-169*.* Oxford, UK: Elsevier.

Almond, P., Kingston, N., Michaels, H., Roeber, E., Warren, S., Winter, P., & Mark, C. (2012). *Technical considerations for developing assessments that include special populations and are based on organized learning models*. Symposium 2011 Topic 3 White Paper. Menlo Park, CA, and Lawrence, KS: SRI International and Center for Educational Testing and Evaluation (CETE).

American Educational Research Association (2011). *Code of ethics of the American Educational Research Association*. http://www.aera.net/Portals/38/docs/About_AERA/CodeOfEthics(1).pdf.

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

American Educational Research Association, American Psychological Association y National Council on Measurement in Education. (2018). *Estándares para pruebas educativas y psicológicas* (M. Lieve, Trans.). American Educational Research Association. https://www.testingstandards.net/uploads/7/6/6/4/76643089/9780935302745_web.pdf.

Andrews-Todd, Mislevy, R. J., LeMar, M., & de Klerk, S. (2021). Virtual performance assessments. In A. von Davier, R. J. Mislevy, & Hao, J. (eds). *Computational psychometrics: New methodologies for a new generation of digital learning and assessment with examples in R and Python*. Springer.

Association of Test Publishers (2002). *ATP Guidelines for Computer-Based Testing*. Author.

Association of Test Publishers (2013). *Assessment Security Options: Considerations by Delivery Channel and Assessment Model*. Author.

Association of Test Publishers (2017). *EU General Data Protection Regulation Compliance Guide*. Author.

Association of Test Publishers (2019). *Privacy in Practice Bulletin #1: Customer Guidance on Privacy Compliance.* ATP.

Association of Test Publishers (2019). *Privacy in Practice Bulletin #2: What is Personal Data in Testing*? Author.

Association of Test Publishers (2019). *Privacy in Practice Bulletin #3: Obligations of Processors in Assessment Services.* Author.

Association of Test Publishers (2019). *Privacy in Practice Bulletin #4: Deletion Request from a Test Taker.* Author.

Association of Test Publishers (2019). *Privacy in Practice Bulletin #5: Breach Management: Step 1-Preparation.* Author.

Association of Test Publishers (2019). *Privacy in Practice Bulletin #6: Preparing for the California Consumer Privacy Act*. Author.

Association of Test Publishers (2020). *Privacy Guidance When Using Video in the Testing Industry*. Author.

Association of Test Publishers (2020). *Privacy in Practice Bulletin #7: Security Standards and the Assessment Industry*. Author.

Association of Test Publishers (2020). *Privacy in Practice Bulletin #8: Privacy by Design and By Default – Demystified*. Author.

Association of Test Publishers (2020). *Privacy in Practice Bulletin #9: Privacy Considerations in Online/Remote Proctoring*. Author.

Association of Test Publishers (2020). *Privacy in Practice Bulletin #10: Breach Management: Step 2 - Investigation and Notification.* Author.

Association of Test Publishers (2020). *Privacy in Practice Bulletin #11: Distinguishing Joint Controller Relationships*. Author.

Association of Test Publishers (2021). *Artificial Intelligence and the Testing Industry: A Primer*. Author.

Association of Test Publishers (August 2021). Before the *European Commission: Comments on Proposed Regulation on the Use of AI*. https://atpu.memberclicks.net/atp-ai-comments.

Association of Test Publishers (2021). *Privacy in Practice Bulletin #12: Complying with California Consumer Privacy Act*. Author.

Association of Test Publishers (January 2022). Artificial Intelligence Principles. https://atpu.memberclicks.net/ai-principles.

Association of Test Publishers (2022). *Privacy in Practice Bulletin #13: Managing a Breach: Step 3 – Review and Remediation*. Author.

Association of Test Publishers & Institute for Credentialing Excellence (2017). *Innovative item types white paper and portfolio*. Authors.

Attali, Y. (2018). *Automatic Item Generation unleashed: An evaluation of a large-scale deployment of item models.* In C. Penstein Rosé et al. (Eds.). (2018). AIED 2018, LNAI 10947, pp. 17–29, Springer International Publishing. https://doi.org/10.1007/978-3-319-93843-1_2

Azano, A. P., & Stewart, T. T. (2015). Exploring place and practicing justice: Preparing preservice teachers for success in rural schools. *Journal of Research in Rural Education*, *30*(*9*), 1-12.

Barnett, T. (2000). Reading "whiteness" in English Studies. *College English, 63*(1) 9 – 37.

Barton, K. (2020). Contextual barriers to validity in adaptive instruction and assessment. In *Adaptive Instructional Systems*. Springer Nature, Switzerland.

Battista, A. Ellenwood, D., Gregory, L., & Higgins, S. (2015). Seeking social justice in the ACRL framework. *Comminfolit*, *9*(*2*), 111-125. https://doi.org/10.15760/comminfolit.2015.9.2.188.

Bender, E., Friedman, B. (2018). Data statements for Natural Language Processing: Toward mitigating system bias and enabling better science. *OpenReview*. Openreview.net/pdf?id=By4oPeX9f

Benitez, M. (2010). Resituating cultural centers within a social justice framework. In L. D. Patton (Ed.). *Culture centers in higher education: Perspectives on identity, theory, and practice* (pp. 119–134). Sterling, VA: Stylus.

Bennett, R. E. (2015). The changing nature of educational assessment. *Review of Research in Education*, *39*(1), 370-407.

Berman, A.I., Haertel, E.H., & Pellegrino, J.W. (2020). (Eds.), *Comparability issues in large-scale assessment: Issues and recommendations*. Washington, DC: National Academy of Education Press.

Bishop, N.S., & Davis-Becker, S. (2016). *Preparing examinees for test taking: Guidelines for test developers*. In S. Lane, M.R. Raymond, & T.M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 554-566). New York, NY: Routledge.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7–74.

Bolsinova, M., & Tijmstra, J. (2018). Improving precision of ability estimation: Getting more from response times. *British Journal of Mathematical and Statistical Psychology*, *71*(1), 13-38.

Bowker, L. (2002). *Computer-aided translation technology: A practical introduction*. Ottawa: University of Ottawa Press.

Brown, M., Dehoney, J., & Millichap, N. (2015). *The next generation digital learning environment*. Boulder, CO: Educause. Retrieved from: https://library.educause.edu/~/media/files/library/2015/4/eli3035-pdf.pdf

Bugbee, A.C. (1996). The equivalence of paper-and-pencil and computer-based testing*. Journal of Research on Computers in Education*, 28(3): 282-299.

Camara, W. J., & Davis, L. (2022). *Fairness concerns resulting from innovations and applications of technology to assessment.* In J. Jonson, K. Geisinger (Eds.). *Fairness issues and solutions in educational and psychological testing: implications for researchers, practitioners, policy makers, and the public* (pp. 163-186). Washington, DC: American Educational Research Association.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Chen, G. (2018, Aug. 6). When teachers cheat: The standardized test controversies. *Public School Review.* Retrieved from https://www.publicschoolreview.com/blog/when-teachers-cheat-the-standardized-test-controversies.

Chiao, C., & Chiu, C.-H. (2018). The mediating effect of ICT usage on the relationship between students' socioeconomic status and achievement. *Asia-Pacific Education Researcher*, *27*(*2*), 109-121.

Cole, B.S., Lima-Walton, E., Brunnert, K., Burt Vesey, W., & Raha, K. (2020). Taming the firehose: Unsupervised machine learning for syntactic partitioning of large volumes of automatically generated items to assist automated test assembly. *Journal of Applied Testing Technology*, Vol 21(1), 1-11.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton-Mifflin.

Council of Chief State School Officers & Association of Test Publishers (2013). *Operational best practices for statewide large-scale assessment programs*. Washington, DC: Council of Chief State School Officers.

Council of Chief State School Officers (2019). *Revising the definition of formative assessment.* Washington, DC: Council of Chief State School Officers. Retrieved from https://ccsso.org/sites/default/files/2018-06/Revising%20the%20Definition%20of%20Formative%20Assessment.pdf.

Crocker, L. (2006). Preparing examinees for test taking: Guidelines for test developers and test users. In S.M. Downing, & T.M. Haladyna (Eds.), *Handbook of test development* (1st ed., pp. 115-128). Mahwah, NJ: Lawrence Erlbaum Associates.

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

DAMA International. (2017). *DAMA-DMBOK Data Management Body of Knowledge*, 2nd Edition. Basking Ridge, NJ: Technics Publications.

De Ayala, R. J., Plake, B. S., & Impara, J. C. (2001). The impact of omitted responses on the accuracy of ability estimation in item response theory*. Journal of Educational Measurement, 38,* 213-234.

Dept, S., Ferrari, A. and Halleux, B. (2017). Translation and cultural appropriateness of survey material in large-scale assessments. In P. Lietz, J. Cresswell, K. Rust and R. Adams (Eds.), *Implementation of large-scale education assessments*. Chichester: John Wiley & Sons, 168-191.

Dolan, R. P., Burling, K., Harms, M., Strain-Seymour, E., Way, W., & Rose, D. H. (2013). *A Universal Design for learning-based framework for designing accessible technology-enhanced assessments* (Research Report). Pearson. https://doi.org/10.13140/RG.2.2.16823.85922.

Dorans, N. J. (2004). Equating, concordance, and expectation. *Applied psychological measurement*, 28(4): 227-246.

Dorans, N.J., Moses, T.P. & Eignor, D. R. (2010). *Principles and practices of test equating*. ETS Research Report (10-29). Princeton, NJ: ETS.

Dorans, N. J., & Puhan, G. (2017). Contributions to score linking theory and practice. In R. E. Bennett & M. von Davier (Eds.), *Advancing human assessment: The methodological, psychological and policy contributions of ETS* (pp. 79–132). Springer Science + Business Media. https://doi.org/10.1007/978-3-319-58689-2_4.

Educational Testing Service (2021). *Best practices for constructed-response scoring*. Princeton, NJ: Author.

Ercikan, K. (1998). Translation effects in international assessments. *International Journal of Educational* Research, 29, 543-553.

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing, 4*, 199-215.

Ercikan, K., Asil, M., & Grover, R. (2018). Digital divide: A critical context for digitally-based assessments. *Education Policy Analysis Archives*, *26*(*51*). http://dx.doi.org/10.14507/epaa.26.3817.

Erl, T. and Mahmood, Z. & Ricardo Puttini, R. (2013) *Cloud Computing: Concepts, technology & architecture*. Prentice Hall, Upper Saddle River, NJ.

European Commission (2019). *Ethics guidelines for trustworthy AI, High-Level expert group on AI*.

European Commission (2020). *Assessment List for trustworthy Artificial Intelligence, high-level expert group on AI.*

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement, 45,* 225-245.

Flaherty, C. (2020, May 11). Big proctor. *Inside Higher Ed*. https://www.insidehighered.com/news/2020/05/11/online-proctoring-surging-during-covid-19.

Folk, V. G., & Smith, R. L. (2002). Models for delivery of CBTs. In C. Mills, M. Potenza, J. Fremer, & W. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 41–66). Mahwah, NJ: Lawrence Erlbaum.

Foster, D. (2016). Testing technology and its effects on test security. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 235–255). New York, NY: Routledge.

Fox, C., & Jones, R. (2019). *The Broadband Imperative III: Driving connectivity, access and student success*. Washington, DC: State Educational Technology Directors Association (SETDA).

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H.M., Daumé, H., & Crawford, K. (2018). *Datasheets for Datasets*. *ArXiv, abs/1803.09010*.

Gibson, W. M., & Weiner, J. A. (1998). Generating random parallel test forms using CTT in a computer-based environment. *Journal of Educational Measurement*, 35, 297–310.

Gierl, M. J., & Lai, H. (2013). Instructional topics in educational measurement (ITEMS) module: Using automated processes to generate test items. *Educational Measurement: Issues and Practice*, *32*(3), 36-50.

Gordon, E.W. (2020). Toward assessment in the service of learning. *Educational Measurement: Issues and Practice*, 39(3), 72–78.

Greenfield, P. M. (1997). You can’t take it with you: Why ability assessments don’t cross cultures. *American Psychologist*, *52*(*10*), 1115-1124.

Grisay, A. (2003). Translation procedures in OECD/PISA 2000 international assessment. *Language Testing,* 20(2), 225-240.

Haladyna, T.M., & Downing, S.M. (2004). Construct-Irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, 23(1), 17-27. https://doi.org/10.1111/j.1745-3992.2004.tb00149.x.

Hambleton, R. K., Merenda, P. F. & Spielberger, C. D. (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hambleton, R. K. & Zenisky, A. L. (2011). Translating and adapting tests for cross-cultural assessments. In D. Matsumoto & F. J. R. van de Vijver (Eds.), *Cross-Cultural Research Methods in Psychology*. Cambridge: Cambridge University Press, pp. 46-74.

Hambleton, R. K. & Zenisky, A. L. (2018). *Score reporting and interpretation*. In W. J. van der Linden (Ed.), *Handbook of Modern Item Response Theory* (2nd ed.).

Hamilton, J., Reddel, S. & Spratt, M. (2001). *Teachers' perceptions of on-line rater training and monitoring*. System. 29. 505-520. https://doi.org/10.1016/S0346-251X(01)00036-7.

Harkness, J. A. (2003) Questionnaire translation. In J. A. Harkness, F.J. R. Van de Vijver, & P. Ph. Mohler (Eds.), *Cross-cultural survey methods* (pp. 35–56). New York: Wiley.

Hayes, A., Turnbull, A., & Moran, N. (2018). Universal design for learning to help all children read: Promoting literacy for learners with disabilities. Washington, D.C.: USAID. Retrieved from https://www.edu-links.org/sites/default/files/media/file/Literacy%20for%20All%20toolkit_v4.1_0.pdf.

Heffernan, K. (2020). Loaded questions: The framework for information literacy through a DEI lens. *College & Research Libraries News*, September 2020, 382-386.

Herold, B. (2020, April 10). The disparities in remote learning under coronavirus (in charts). *Education Week*. https://www.edweek.org/ew/articles/2020/04/10/the-disparities-in-remote-learning-under-coronavirus.html.

Hind, M., Mehta, S., Mojsilovic, A., Nair, R., Ramamurthy, K.N., Olteanu, A., & Varshney, K.R. (2018). Increasing Trust in AI Services through Supplier's Declarations of Conformity. *IBM Journal of Research and Development, 63*, 6:1-6:13.

Howard, A., & Borenstein, J. (2018). The ugly truth about ourselves and our robot creations: The problem of bias and social inequity. *Science & Engineering Ethics*, *24*, 1521-1536.

Huff, K. L, & Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practice, 20(3)*, 16-25.

Iliescu, D. (2017). *Adapting tests in linguistic and cultural situations*. New York: Cambridge University.

International Organization for Standardization and International Electrotechnical Commission (2013). ISO/IEC Standard 27001:2013 Information technology – Security techniques – Information security management systems – Requirements.

International Organization for Standardization (2013). ISO 27001: 2013, *Information Technology – Security techniques – Information security management systems – Requirements*.

International Organization for Standardization (2013). ISO 27002:2013*, Information technology – Security techniques – Code of practice for information controls.*

International Organization for Standardization (2013). *ISO 8000 (2020)*. *Data quality: Overview.*

International Organization for Standardization (2019). ISO 27001, Information Technology – Security techniques. https://www.iso.org/standard/54534.html.

International Test Commission (2005). *Guidelines for computer-based and internet delivered testing*. International Test Commission.

International Test Commission. (2012). *International guidelines on quality control in scoring, test analysis, and reporting of test scores*. https://www.intestcom.org/page/17.

International Test Commission (2013a). I*TC guidelines on quality control in scoring, test analysis, and reporting of test scores*. International Test Commission. Available at https://www.intestcom.org/files/guideline_quality_control.pdf.

International Test Commission (2013b). *ITC guidelines on test use*. International Test Commission. Available at https://www.intestcom.org/files/guideline_test_use.pdf.

International Test Commission (July 2014). *The ITC guidelines on the security of tests, examinations and other assessments*. https://www.intestcom.org/files/guideline_test_security.pdf

International Test Commission (2018a). *ITC guidelines for the large-scale assessment of linguistically and culturally diverse populations*. International Test Commission. Available at https://www.intestcom.org/files/guideline_diverse_populations.pdf.

International Test Commission (2018b). ITC guidelines for translating and adapting tests (2nd Edition). *International Journal of Testing*, *18, 2*, 101-134. DOI: 10.1080/15305058.2017.1398166. https://doi.org/10.1080/15305058.2017.1398166

John, T., & Misra P. (2017) *Data lake for enterprises*. Packt Publishing Ltd. Birmingham, UK.

Juran, J. M., Gryna, F. M., Godfrey, A. B., Schilling, E. G., Hoogstoel, R. E., & Joseph, J. (1999). *Juran's quality handbook*. McGraw Hill. https://books.google.com/books?id=beVTAAAAMAAJ

Kane, M. (1982). A sampling model for validity. *Applied Psychological Measurement*, *6* (2), 125-160. https://doi.org/10.1177/014662168200600201.

Kane, M. (2006). Validation. In R. L. Brennan (Ed). *Educational measurement* (4th edition, pp. 17-64). Washington, DC: American Council on Education/Praeger.

Kane, M. (2011). The errors of our ways. *Journal of Educational Measurement*, *48* (1), 12-30. https://doi.org/10.1111/j.1745-3984.2010.00128.x.

Kane, M. (2013). Validating interpretations and uses of test scores. *Journal of Educational Measurement*, *50* (1), 1-73. https://doi.org/10.1111/jedm.12000.

Katz, V. S., & Gonzalez, C. (2015). Community variations in low-income Latino families' technology adoption and integration. *American Behavioral Scientist*, *60*(*1*), 59-80.

Kemp, N. & Grieve, R. (2014). Face-to-face or face-to-screen? Undergraduates' opinions and test performance in classroom vs. online learning. *Frontiers in Psychology,* https://www.frontiersin.org/article/10.3389/fpsyg.2014.01278.

Ketterlin-Geller, L. R. (2008). Testing students with special needs: A model for understanding the interaction between assessment and student characteristics in a universally designed environment. *Educational Measurement: Issues and Practice, 27*(3), 3–16.

Ketterlin-Geller, L.R., Johnstone, C.J., & Thurlow, M.L. (2015). Universal design in assessment. In Burgstahler, S.E. (Ed.), *Universal design in higher education: From principles to practice* (2nd ed.), pp. 163-175. Cambridge, MA: Harvard Education Press.

Kingsbury, G.G., & Weiss, D.J. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, 21(4), 361–375.

Klinkenberg, S., Straatemeier, M., & van der Maas, H. L. (2011). Computer adaptive practice of math ability using a new item response model for on-the-fly ability and difficulty estimation. *Computers & Education*, *57*(2), 1813-1824.

Klosowski, T. (July 2020). *Facial Recognition is everywhere: Here's what we can do about it.* New York Times. https://www.nytimes.com/wirecutter/blog/how-facial-recognition-works/.

Kolen, M.J. & Brennan, R.L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). Springer Science + Business Media.

Kunin, M., Julliard K., & Rodriguez T. (2014). Comparing face-to-face, synchronous, and asynchronous learning: postgraduate dental resident preferences. *J Dent Educ. 78(6)*, 856-66. PMID: 24882771.

Lai, E.R., & Waltman, K. (2008). Test preparation: Examining teacher perceptions and practices. *Educational Measurement: Issues and Practices, 27*, (2), 28-45.

Lane, S. (2014). Validity evidence based on testing consequences. *Psicothema*. doi:10.7334/psicothema2013.258.

Langenfield, T. (2020). Internet-based proctored assessment: Security and fairness issues. *Educational Measurement: Issues and Practice*, 39(3), 24-27.

Lee, D., Buzick, H., Sireci, S. G., Lee, M., & Laitusis, C. (2021). Embedded accommodation and accessibility support usage on a computer-based statewide achievement test. *Practical Assessment, Research & Evaluation*, *26*(25). Available online: https://scholarworks.umass edu/pare/vol26/iss1/25/.

Llabre, M. M., Clements, N. E., Fitzhugh, K. B., Lancelotta, G. (1987). The effect of computer-administered testing on test anxiety and performance. *Journal of Educational Computing Research, 3*(4), 429–433. Retrieved from https://psycnet.apa.org/record/1989-10285-001.

Lottridge, S., Nicewander, A., Schulz, M., & Mitzel, H. (2010). Comparability of paper-based and computer-based tests: A review of the methodology. In P. C. Winter (Ed). *Evaluating the comparability of scores from achievement test variations* (pp. 119-152). Council of Chief State School Officers: Washington, D. C.

Lubin, G. (2013, Oct. 8). Here's the stupid way someone got caught cheating on the bar exam. *Business Insider.* Retrieved from https://www.businessinsider.com/cheating-on-the-bar-exam-2013-10.

Luecht, R.M. (2005). Some useful cost-benefit criteria for evaluating computer-based test delivery models and systems. *Journal of Applied Testing Technology*, 7(2).

Luecht, R. M. (2016). Computer-based test delivery models, data, and operational implementation issues. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 179–205). New York, NY: Routledge.

Luecht, R.; & Burke, M. (2020). Reconceptualizing items: From clones and automatic item generation to task model families. In R. Lissitz & H. Jiao (Eds.), *Applications of artificial intelligence to assessment*. Baltimore, MD: Information Age Publishers.

Luecht, R, & Sireci, S. (2011). *A review of models for computer-based testing*. ETS Research Report 2011-12. Princeton, NJ: ETS.

Maris, G., & van der Maas, H. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika*, *77*(4), 615-633.

Martineau, J., & Dadey, N. (2016). *Final report on online interruptions of the Spring 2015 Smarter Balanced Assessment Administration in Montana, Nevada, and North Dakota.* Dover NH: The National Center for the Improvement of Educational Assessment.

Martineau, J., Domaleski, C., Egan, K., Patelis, T., & Dadey, N. (2015, November). *Recommendations for addressing the impact of test administration interruptions and irregularities*. Washington, DC: Council of Chief State School Officers (CCSSO). Available online: https://www.nciea.org/library/recommendations-addressing-impact-test-administration-interruptions-and-irregularities.

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessments, and instruction, *Review of Educational Research 4*, 1332-1361.

Mayson, S. G. (2019). Bias in, bias out. *Yale Law Journal*, *128*, 2218-2300.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York, NY: MacMillan.

Mislevy, R. J., Behrens, J. T., DiCerbo, K. E., Frezzo, D. C., & West, P. (2012). Three things game designers need to know about assessment. In *Assessment in game-based learning* (pp. 59-81). Springer, New York, NY.

Mislevy, R.J., Corrigan, S., Oranjc, A. DiCerbo, K., Bauer, M.I., Davier, A., John, M. (2016). Psychometrics and game-based assessment. In F. Drasgow (Ed.). *Technology and testing: Improving educational and psychological measurement* (pp. 23-48). New York, NY: Routledge.

Molenaar, D., Tuerlinckx, F., & van der Maas, H.L.J. (2015). A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times. *Multivariate Behavioral Research, 50,* 56-74.

Moore, R., Vitale, D., Stawinoga, N. (2018). *The Digital divide and educational equity: A Look at students with very limited access to electronic devices at home*. ACT Center for Equity in Learning. Retrieved September 15, 2020, from https://www.act.org/content/dam/act/unsecured/documents/R1698-digital-divide-2018-08.pdf.

National Academies of Sciences, Engineering, and Medicine (2019). *Monitoring educational equity*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25389.

National Center on Educational Outcomes (2015). *Considerations when including students with disabilities in test security policies*. Policy Directions, No. 23. https://nceo.umn.edu/docs/OnlinePubs/Policy23/PolicyDirections23.pdf.

National Center on Educational Outcomes. (2022). *Universally designed assessments: Better tests for everyone!* Retrieved from https://nceo.info/Resources/publications/onlinepubs/Policy14.htm

National Conference of State Legislatures (2021). https://www.ncsl.org/research/telecommunications-and-information-technology/2020-legislation-related-to-artificial-intelligence.aspx.

National Council on Measurement in Education (2012). *Testing and data integrity in the administration of statewide student assessment programs*. Madison, WI: Author.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment.* Committee on the Foundations of Assessment. Pelligrino, J., Chudowsky, N., & Glaser, R., editors. Board on Testing and Assessment, Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

Newton, P. E. (2010). Contrasting conceptions of comparability. *Research Papers in Education*, 25(3), 285-292. https://dx.doi.org/10.1080/02671522.2010.498144.

OECD (2013). *Recommendation of the Council concerning guidelines governing the protection of privacy and transborder flows of personal data.* (“OECD Privacy Principles”) https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0188.

OECD (2019, May). *Principles on Artificial Intelligence*. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

OECD (2020), *PISA 2018 Results (Volume V): Effective policies, successful schools*, PISA, OECD Publishing, Paris, https://doi.org/10.1787/ca768d40-en.

OECD (2021), 21st-Century Readers: Developing literacy skills in a digital world, PISA, OECD Publishing, Paris, https://doi.org/10.1787/a83d84cb-en.

Oswald, F. (2020). Future research directions for big data in psychology. In S. E. Woo, L. Tay, & R. Proctor (Eds.). *Big data in psychological research* (pp. 427-441). Washington, DC: APA.

Padilla, J., & Benítez, I. (2014). Validity evidence based on response processes. *Psyicothema*, 26, 136-144.

Paris, B. (2020, Jan. 21). The next cheating scandal. *Inside Higher Education.* Retrieved from https://www.insidehighered.com/admissions/views/2020/01/21/next-testing-scandal-could-come-any-number-directions-opinion.

Parker, H. E. (2015). *Digital badges to assess bloom's affective domain*. The National Teaching & Learning Forum, 24(4), 9–11.

Parshall, C. G., Spray, J. A., Kalohn, J., & Davey, T. (2002). *Practical considerations in computer-based testing.* New York, NY: Springer Verlag.

Popham, W. J. (1991). Appropriateness of teachers' test preparation practices. *Educational Measurement: Issues and Practices, 10*, 12-15.

Popham, W. J. (2003). *Test better, teach better: The instructional role of assessment.* Alexandria, VA: Association for Supervision and Curriculum.

Powers, J. R., Musgrove, A. T., & Nichols, B. H. (2020). Teachers bridging the digital divide in rural schools with 1:1 computing. *Rural Educator*, *41*(*1*), 61-76.

Prometric, Inc. (2020). *Cheating 2.0: How to fight back against cheaters.* Retrieved from https://www.prometric.com/test-owners/resources/cheating-20.

Psaltis, A. (2017) Streaming data: Understanding the real-time pipeline. Manning Publications, New York.

Ranger, J. (2013). A note on the hierarchical model for responses and response times in tests of van der Linden (2007). *Psychometrika*, *78*(3), 538-544.

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC [General Data Protection Regulation (GDPR)], OJ 2016 L 119/1. EUR-Lex - 32016R0679 - EN - EUR-Lex (europa.eu).

Randall, J. (2021). "Color-Neutral" is not a thing: Redefining construct definition and representation through a justice-oriented critical antiracist lens. *Educational Measurement: Issues and Practice*. https://doi.org/10.1111/emip.12429.

Rose, D. H., Meyer, A., & Hitchcock, C. (2005). *The universally designed classroom: Accessible curriculum and digital technologies.* Cambridge, MA: Harvard Education Press.

Roturier, J. (2019). XML for translation technology. In M. O'Hagan (Ed.), *The Routledge Handbook of Translation and Technology*. Routledge, London.

Russel, M., Goldberg, A., & O'Connor, K. (2003). Computer-based testing and validity: A look back and into the future. *Assessment in Education: Principles, Policy, and Practice, 10,* 279-293. Retrieved from https://www.tandfonline.com/doi/abs/10.1080/0969594032000148145.

Salen, K., & Zimmerman, E. (2004). *Rules of Play: Game design fundamentals*. Cambridge: MIT Press.

Schmidt, F. L. (1988). Validity generalization and the future of criterion-related validity. In H. Wainer & H.I. Braun (Eds.), *Test validity* (pp. 173-189). Hillsdale, New Jersey: Lawrence Erlbaum.

Sensoy, O. & DiAngelo, R. (2012). *Is everyone really equal? An introduction to key concepts in social justice education, 1^st edition*. Teacher's College Press: New York.

Settles, B., LaFlair, G. T., & Hagiwara, M. (2020). Machine-driven language assessment. Transactions of the Association for Computational Linguistics, 8: 247-263. https://doi.org/10.1162/tac1_a_00310.

Sireci, S. G. (1997). Problems and issues in linking tests across languages. *Educational Measurement: Issues and Practice, 16(1)*, 12-19.

Sireci, S. G. (2005). Unlabeling the disabled: A perspective on flagging scores from accommodated test administrations. *Educational Researcher*, *34*(1), 3-12.

Sireci, S. G. (2020). Standardization and UNDERSTANDardization in educational assessment. *Educational Measurement: Issues and Practices*, *39* (3), 100-105. https://doi.org/10.1111/emip.12377.

Sireci, S. G., & Faulkner-Bond (2014). Validity evidence based on test content. *Psicothema*, *26*, 100-107. doi: 10.7334/psicothema2013.256.

Sireci, S. G., & O'Riordan, M. (2020). Comparability issues in assessing individuals with disabilities. In A.I. Berman, E.H. Haertel, & J.W. Pellegrino (Eds.), *Comparability Issues in Large-Scale Assessment: Issues and recommendations* (p. 117-204). Washington, DC: National Academy of Education Press.

Sireci, S.G., Patsula, L., & Hambleton, R. K. (2005). Statistical methods for identifying flawed items in the test adaptations process. In R.K. Hambleton, P. Merenda, & C. Spielberger (Eds.). *Adapting educational and psychological tests for cross-cultural assessment* (pp. 93-115). Hillsdale, NJ: Lawrence Erlbaum.

Sireci, S. G., & Randall, J. (2021). Evolving notions of fairness in testing in the United States. In M. Bunch & B. Clauser (Eds.), *The history of educational measurement: Key advancements in theory, policy, and practice* (pp. 111-135). New York: Routledge.

Sireci, S. G., Rios, J. A., & Powers, S. (2016). Comparing test scores from tests administered in different languages. In N. Dorans & L. Cook (Eds.) *Fairness in educational assessment and measurement* (pp. 181-202). New York: Routledge.

Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representations. In S. M. Downing & T. M. Haladyna, (Eds.), *Handbook of test development* (pp. 329-347). Mahwah, NJ: Lawrence Erlbaum Associates.

Sireci, S. G., & Zenisky, A. L. (2016). Computerized innovative item formats: Achievement and Credentialing. In S. Lane, T. Haladyna, & M. Raymond (Eds.). *Handbook of test development* (pp. 313-334). Washington, DC: National Council on Measurement in Education.

Society for Personality Assessment (2022). *Tele-Assessment of Personality and Psychopathology COVID-19 Task Force to Support Personality Assessment*. https://assets.noviams.com/novi-file-uploads/spa/PDFs__Documents/COVID-19_Resources/Tele-Assessment_Resources/SPA_Personality_Tele-Assessment-Guidance_6_10_20.pdf.

Sottilare, R. A., & Schwarz, J. (Eds.). (2020). *Adaptive instructional systems: Second International Conference,* AIS 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings (Vol. 12214). Springer International Publishing. https://doi.org/10.1007/978-3-030-50788-6.

Stocking, M.L., Smith, R., & Swanson, L. (2000*). An investigation of approaches to computerizing the GRE subject tests.* Research Report 00-04. Princeton, NJ: Educational Testing Service.

Story, M., Mueller, J., & Mace, R. (1998). *The Universal Design File: Designing for people of all ages and abilities.* Revised Edition. Center for Universal Design, NC State University. ERIC Number: ED460554.

Swauger, S. (2020). Our bodies encoded: Algorithmic test proctoring in higher education. *Hybrid Pedagogy*. Retrieved September 15, 2020, from https://hybridpedagogy.org/our-bodies-encoded-algorithmic-test-proctoring-in-higher-education/.

Sweet, E. (2016). *Data lineage and compliance*. (Retrieved November 24, 2020, from https://www.isaca.org/resources/isaca-journal/issues/2016/volume-5/data-lineage-and-compliance).

Thompson, S.J., Thurlow, M.L., & Malouf, D. (2004, May). Creating better tests for everyone through universally designed assessments. *Journal of Applied Testing Technology, 10*(2). See *https://www.testpublishers.org/journal-of-applied-testing-technology*

Tippins, N.T., Oswald, F.L., & McPhail, S.M. (2021). Scientific, legal, and ethical concerns about AI-based personnel selection tools: A call to action. *Personnel Assessment and Decisions*. No. 7 (2), 1-22.

UK Information Commissioner's Office (2020). Guide to the General Data Protection Regulation (GDPR): https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287.

van der Linden, W. J., Klein Entink, R. H., & Fox, J. P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement*, *34*(5), 327-347.

van Rijn, P. W., & Ali, U. S. (2018). A generalized speed-accuracy response model for dichotomous items. *Psychometrika*, *83*(1), 109-131.

von Braun, J., & Gatzweiler, F. W. (2014). Marginality: An overview and implications for policy. In J. von Braun & F. W. Gatzweiler (Eds.). *Marginality: Addressing the nexus of poverty, exclusion and ecology*. New York: Springer.

von Davier, A. A., Deonovici, B., Yudelson, M., Polyak, S. T., & Woo, A. (2019). Computational psychometrics approach to holistic learning and assessment systems. *Frontiers in Education*, 4 (69), doi: 10.3389/feduc.2019.00069.

von Davier A.A., Zhu, M., & Kyllonen, P.C. (Eds). (2017). *Innovative assessment of collaboration.* Cham, Switzerland: Springer International Publishing.

Wang., T. & Kolen, M. J. (2001). Evaluating comparability in computer adaptive testing: Issue, criteria and an example. *Journal of Educational Measurement*, 38(1): 19-49.

Ward, T. J., Hooper, S. R., & Hannafin, K. M. (1989). The effect of computerized tests on performance and attitudes in college students. *Journal of Educational Computing Research, 5*, 327-333.

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education, 20*(1), 7-25.

Weiner, J. A., & Foster, D. (2018). *Licensing and Certification*, Ch. 2., In Scott, Bartram, & Reynolds, Eds., *Next Generation Technology-Enhanced Assessment*. New York: Cambridge University Press.

Wells. C. S. (2021). *Assessing measurement invariance for applied research*. Cambridge: Cambridge University Press.

Williamson, D., Xi, X., & Breyer, F. J. (2012). A framework for the evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31(1),* 2-13.

Winter, P.C. (2010). *Evaluating the comparability of scores from achievement test variations*. Council of Chief State School Officers. Washington DC: CCSSO.

Wise, S. L. (2015). Effort Analysis: Individual Score Validation of Achievement Test Data. *Applied Measurement in Education*, *28*(3), 237–252. https://doi.org/10.1080/08957347.2015.1042155.

Wollack, J. A., & Fremer, J. J. (2013). *Handbook of test security*. New York: Routledge.

Wools, S., Molenaar, M., Hopster-den Otter, D. (2019). *The validity of technology enhanced assessments – threats and opportunities*. In B. P. Veldkamp & C. Sluijter (Eds.), *Theoretical and practical advances in computer-based educational measurement* (pp. 3-19). Cham, Switzerland: Springer International Publishing.

Wright, A. J., Mihura, J. L., Pade, H., & McCord, D. M. (2020). *Guidance on psychological tele-assessment during the COVID-19 crisis*. https://www.apaservices.org/practice/reimbursement/health-codes/testing/tele-assessment-covid-19.

Yaneva, V., Ha, L., Baldwin, P., & Mee, J. (2020). *Predicting item survival for multiple choice questions in a high-stakes medical exam*. Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pp. 6812–6818, Marseille, 11–16 May 2020.

Yannakoudakis, H. & Cummins, R. (2015). *Evaluating the performance of Automated Text Scoring systems*. In Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 213–223, Denver, Colorado: Association for Computational Linguistics.

Zenisky, A. L., & Hambleton, R. K. (2013). From "Here's the Story" to "You're in Charge": Developing and maintaining large-scale online test and score reporting resources. In M. Simon, M. Rousseau, & K. Ercikan (Eds.), *Improving large-scale assessment in education* (pp.175-185). New York, NY: Routledge.

Zenisky, A. L., & Hambleton, R. K. (2015). Test score reporting: Best practices and issues. In S. Lane, M. Raymond, and T. Haladyna (Eds.), *Handbook of test development* (2nd ed.), p. 585-602. New York, NY: Routledge.

Zickar, M. J. (2018). *Using social media for assessment*, Ch. 8., In Scott, Bartram, & Reynolds, Eds., *Next Generation Technology-Enhanced Assessment*. New York: Cambridge University Press.

Zumbo, B. D., & Hubley, A. M. (Eds.) (2017). *Understanding and investigating response processes in validation research*. Cham, Switzerland: Springer Press.

Zwick, R. (2006). Higher education admissions testing. In R.L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 647-679). Westport, CT: Praeger.

Appendix: Locked-down Browser Checklist

Test sponsors may use this checklist to verify functionality of a testing vendor’s locked-down browser.

Prevent access to non-authorized tools

- □ Display test full screen
- □ Block virtual machines
- □ Block remote desktop
- □ Block applications
- □ Block unauthorized websites
- □ Provide secure method to verify locked-down browser is running
- □ Block multi-monitors and lock screens from being used to cheat

Support remote proctoring

- □ Detect virtual video, virtual microphones, and duplicate input devices
- □ Prevent a test from being delivered if external remote proctoring software ceases to run

Prevent content from being stolen or exposed

- □ Block screen captures
- □ Clear cut, copy, and paste buffers
- □ Support clearing cache before and after testing
- □ Block proxy server attacks that bypass HTTPS protections
- □ Block printing

Required features

- □ Block assistive technologies not related to accessibility
- □ Block content on additional monitors
- □ Block using lock screens to show custom images
- □ Block gestures that allow access to content
- □ Upload security issues such as blocked processes or invalid key attempts
- □ Support automatic software updates
- □ Support automatic configuration updates before each test

Optional features

- □ Support suppressing authorization requests for microphones

Platforms

- □ Support all major platforms
- □ Configurable rendering engines

Privacy

- □ Support uninstall
- □ Limit tracking
- □ Disclosure of all information captured accessibility
- □ Compatibility with assistive software for accessibility