

Comparisons of and Concerns about Two Testing Application Chapters in the 2014 Standards for Educational and Psychological Testing

Nurliyana Bukhari

School of Education and Modern Languages, Universiti Utara Malaysia, Malaysia

Received June 20, 2020; Revised August 3, 2020; Accepted August 25, 2020

Cite This Paper in the following Citation Styles

(a): [1] Nurliyana Bukhari, "Comparisons of and Concerns about Two Testing Application Chapters in the 2014 Standards for Educational and Psychological Testing," *Universal Journal of Educational Research*, Vol. 8, No. 10, pp. 4603-4609, 2020. DOI: 10.13189/ujer.2020.081028.

(b): Bukhari, N. (2020). *Comparisons of and Concerns about Two Testing Application Chapters in the 2014 Standards for Educational and Psychological Testing*. *Universal Journal of Educational Research*, 8(10), 4603-4609. DOI: 10.13189/ujer.2020.081028.

Copyright©2020 by authors, all rights reserved. Authors agree that this article remains permanently open access under the terms of the Creative Commons Attribution License 4.0 International License

Abstract This paper compares two "Testing Application" chapters of the 2014 Standards for Educational and Psychological Testing (hereinafter, *Standards*), jointly published by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). The two chapters are Chapters 10 and 12 of the *Standards*: "Psychological Testing and Assessment (PTA)" and "Educational Testing and Assessment (ETA)". It specifically aims to raise some overarching issues related to these chapters. An in-depth comparative analysis was conducted based on specific similarities of and differences between these two chapters. Both PTA and ETA cover the background of and standards regarding test administration, score interpretation, and the use of scores. However, PTA focuses more on test selection and test security, whereas ETA covers more on the test design and development. The overarching issues, questions, and concerns related to both chapters are discussed along with the results of the analysis. The paper concludes with a description of the differences between the two chapters from the current 2014 *Standards* and those of the previous 1999 version along with some plausible explanations for such discrepancies. The summary and analysis may be useful to test users and graduate students from the psychology and education fields whose interests revolve around testing and assessment practices.

Keywords International Testing Standards, Educational Testing and Assessment, Psychological Testing and Assessment, Testing Applications

1. Introduction

The 2014 *Standards* is a joint publication of three sponsoring associations, the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). It is published to provide updates to the documentation in the previous 1999 edition. Linn (2006) noted that the 1999 version was "widely recognized as the most authoritative statement of professional consensus regarding the development and evaluation of educational and psychological tests" (p.27). Hence, the 2014 publication primarily serves two purposes (AERA et al., 2014, p.1): "[T]o provide criteria for the development and evaluation of tests and testing practices," and "to provide guidelines for assessing the validity of interpretations of test scores for the intended test uses."

The *Standards* (AERA et al., 2014) is intended for professionals who specify, develop, or select tests and for those who interpret or evaluate the technical quality of test results. Other audiences include professional test sponsors,

publishers of tests, and test users. Previous six documents (i.e., series) of the *Standards* (AERA et al., 2014, p. vii) include the Technical Recommendations for Psychological Tests and Diagnostic Techniques (APA, 1954), the Technical Recommendations for Achievement Tests (AERA & NCME, 1955), and the previous versions of the *Standards* (AERA et al., 1966, 1974, 1985, 1999).

In the 2014 *Standards* (AERA et al., 2014), a total of thirteen chapters are categorized into three major parts: Foundations, Operations, and Testing Applications; A summary of the content for the *Standards* (AERA et al., 2014) is provided in Figure 1.

In this paper, two Testing Application chapters of the 2014 *Standards* (AERA et al., 2014) are compared. The two chapters are Chapter 10: The Psychological Testing and Assessment (PTA) chapter and Chapter 12: The Educational Testing and Assessment (ETA) chapter. An in-depth comparative analysis was conducted based on similarities of and differences between these two chapters. The paper also discusses overarching issues, questions, and concerns related to both chapters along with the results of the analysis. It concludes with a description of the differences between the two chapters from the current 2014 *Standards* and those of the previous 1999 version along with some plausible explanations for such discrepancies.

PART I: FOUNDATIONS	PART II: OPERATIONS	PART III: TESTING APPLICATIONS
1. Validity 2. Reliability/Precision and Errors of Measurement 3. Fairness	4. Test Design & Development 5. Scales, Norms, Score Linking, and Cut Scores 6. Test Administration, Scoring, Reporting, & Interpretation 7. Supporting Documentation for Tests 8. Test Takers' Rights and Responsibilities 9. Test Users' Rights and Responsibilities	10. Psychological Testing and Assessment 11. Workplace Testing and Credentialing 12. Educational Testing and Assessment 13. Uses of Tests for Program Evaluation, Policy Studies, and Accountability

Figure 1. The Content Summary of the *Standards* (AERA, APA, & NCME, 2014)

Table 1. Summary of the Standards from Chapter 10 and Chapter 12 of the *Standards* (AERA, APA, & NCME, 2014).

CHAPTER 10: PSYCHOLOGICAL TESTING & ASSESSMENT	CHAPTER 12: EDUCATIONAL TESTING & ASSESSMENT
Cluster 1: Test user qualifications	Cluster 1: Test design & development
10.1 Those who use and interpret psychological tests should narrow their testing and assessment activities to their areas of competence (12.1 in 1999 – with more examples elaborated in the commentary note)	12.1 Test mandates describe how test results are used, identify intended/unintended test impact, and minimize negative consequences (13.1 in 1999)
10.2 Those who select tests and draw inferences from test scores should be familiar with the validity and reliability for the intended uses of the test scores and assessments (12.13 in 1999)	12.2 For a test with multiple purposes, evidence for tests quality (scores validity/reliability) and fairness for each intended use need to be provided (13.2 in 1999)
10.3 Professionals should supervise “interns” closely and make sure they have appropriate knowledge and skills to administer and score tests (combination of 12.8 & 12.9 in 1999 – standard was succinctly reduced and commentary note was added in 2014)	12.3 Everyone (all relevant subgroups in the target population) should have an equal chance to demonstrate proficiency on a test through consideration of all relevant steps of the testing process (e.g., use of technology, administration, scoring, adaptations/accommodations/ modifications) (new)

Table 1 Continuous

Cluster 2: Test selection	12.4 Evidence of test and knowledge/target domain relationship should be provided (13.3 in 1999)
10.4 Select combination of tests that are appropriate for purposes of the assessment (12.5 in 1999)	12.5 Local norms might be useful in conjunction with broad representative norms (published norms) (13.4 in 1999)
10.5 Tests should be suitable for the characteristics and background of the test taker. Normative or validity studies must be available for relevant population before conclusions are made (12.3 in 1999 – commentary note is extended in 2014)	12.6 Computer tests should be held to the same requirements of technical quality (design, models, scoring algorithms) as other tests (13.18 in 1999)
10.6 Select tests that have adequate information (mean difference, statistical significance, effect size/ variance explained by predictor) to sufficiently validate the scores and “end” interpretation of diagnosis if differential diagnosis is needed (12.6 in 1999)	Cluster 2: Test use and interpretation
Cluster 3: Test administration	12.7 Over preparation and over exposure of test prep/ test materials can jeopardize the interpretation of scores (13.11 in 1999)
10.7 Directions and information on test for “how and why” should be provided to test takers to minimize errors (12.10 in 1999 - standard was succinctly reduced in 2014)	12.8 Evidence showing that students had an opportunity to learn the skills and content measured by the test should be provided for high stake tests (e.g., promotion, graduation) (13.5 in 1999)
10.8 Professionals and test administration should follow standard administration procedures and provide settings for testing that facilitate the performance of test takers (12.12 in 1999 – revised to generalize)	12.9 Multiple opportunities (alternative approaches with the same standards of technical quality) should be available for students to demonstrate mastery of skills/knowledge for graduation purpose (13.6 in 1999 - slightly paraphrased)
10.9 Quality control is ensured by taking into account the construct being measured, the capabilities of the test taker, and the purpose of the assessment when deciding to use technology based administration	12.10 Collateral information in educational settings should also be considered before major impact decisions are made (13.7 in 1999 - slightly paraphrased)
Cluster 4: Test interpretation	12.11 Difference or growth scores should be clearly defined, evidence of their validity, reliability/precision (standard error), and fairness should be reported, and complex statistical models should be described and justified (13.17 in 1999 - reworded and comments are elaborated in 2014)
10.10 Keep parties with vested interests out of the test selection and interpretation process to disable inappropriate influences on the test interpretation (12.2 in 1999)	12.12 When comparing scores/subscores from different tests/ test components, overlapping of tests constructs, and reliability (standard error) of the difference scores should be considered. Also, tests should be placed on a common scale and normed on the same population at about the same time (13.8 in 1999)
10.11 Share test scores with test takers when appropriate or when required by law (explicit statement on the enforcement of litigation) (12.20 in 1999 – legal enforcement was explicitly emphasized in 2014 but only implied in 1999)	12.13 Empirical evidence documenting relationship among test scores, instructional programs, and student outcome should be provided when test scores are used to make decisions, implement programs/services, etc. about students with disabilities or ELL. If not available, test users should weight the test results given relevant information about the students (13.9 in 1999 - slightly paraphrased. ELL is not explicitly stated)
10.12 Explanation of other factors that may have influences on test scores and whether they may introduce construct-irrelevant variance. Alternative explanations could also be provided by an analysis of stylistic or other qualitative features of test-taking behavior (combination of 12.14 & 12.19 in 1999 – most statements of standards were combined into commentary note in 2014)	12.14 Supervisors of testing practices should be familiar with evidence for reliability/precision, validity of intended interpretations, and the fairness of test scores and should be able to train others involved in testing practices (qualified professionals) (13.12 in 1999 - comment was not provided)
10.13 Diagnosis should be identified when test scores support it (12.7 in 1999 – commentary note added in 2014)	12.15 Responsible parties should take steps to verify that school personnel who interpret the test results to guide instruction and make decisions within the school context are qualified to do so or are assisted by and consult qualified professionals (13.13 in 1999)

Table 1 Continuous

10.14 Score interpretations should not imply evidence for a relationship (criterion-related evidence) unless the evidence is available (combination of 12.16 (became commentary note in 2014) & 12.17 (the stated standard in 2014) in 1999)	Cluster 3: Test administration, scoring, and reporting
10.15 Interpretations for diagnosis should be supported by collateral information (12.18 in 1999)	12.16 Individuals responsible in test administration and scoring should receive appropriate training/ skills and should adhere to the directions (e.g., test security protocols, accommodations/modifications) provided by test developer (13.10 in 1999 - comment was not provided)
10.16 Evidence for the packaging/combination of tests should be provided; separate evidence is insufficient (combination of 12.4 & 12.5 in 1999)	12.17 Test users should consider appropriate contextual information (different demographics) and be cautioned against misinterpretation. These are meant to ensure meaningful interpretation of the differences (combination of 13.16 and 13.15 in 1999 - reworded and specific details were generalized in 2014)
10.17 Quality of validity evidence should suffice And be verified for computer-generated interpretations of test data (12.15 in 1999 – specific and confusing examples were revised and generalize)	12.18 Score reports should include: clear presentation of the scores, guidelines on how to interpret them, their degree of measurement error, dates of test administration, and relevant norming studies (13.19 in 1999 - reworded and specific details were generalized in 2014)
Cluster 5: Test security	12.19 Evidence should be provided to support any pedagogical recommendation included in the score reports (new in 2014)
10.18 All involved should maintain confidentiality of test materials and test scores (12.11 in 1999 – commentary note was extended in 2014)	

2. Summary of the Two Chapters

PTA and ETA share several similarities in terms of their standards and their background information. Both PTA and ETA cover the background of and standards regarding test administration, score interpretation, and the use of scores (p. 152-168, p. 183-200). However, they differ in the emphasis placed on certain themes. PTA focuses more on test selection (p. 152) and test security (p.168), whereas ETA covers more the test design and development (p. 184).

Chapter 10 of the *Standards* (AERA et al., 2014) focuses on the standards for PTA. The chapter is based on the premise that a psychology test (Bornstein, 2017; Camara, Nathan, & Puente, 2000; Groth-Marnat & Wright, 2016) will meet its intended purpose(s) through meaningful interaction between the stakeholders involved in a setting in which the test is used and a relevant set of sophisticated professional activities to produce useful results, inventories, and interpretations. Chapter 12 delineates standards in the context of ETA. The purpose of the standards and the target audience to whom the standards are currently applied are stated at the beginning of each chapter. The content of the standards from both chapters is summarized and listed in Table 1.

2.1. Specific Similarities of the Two Chapters

Although PTA and ETA are conducted in separate

fields, they share similar test and assessment guidelines with respect to testing practices. In this section, a few key standards from both fields will be highlighted and discussed.

Standards 10.1 and 10.3 from PTA are closely related to Standard 12.15 from ETA. The two PTA standards suggest that professionals should receive proper training and acquire the certification required to administer and interpret the tests. In the ETA chapter, Standard 12.15 states the importance of considering the professional qualifications of the school personnel responsible for interpreting the test scores for the purpose of informing instruction and making decisions within the school context. These standards speak to the appropriateness of the skills and knowledge of professionals and supervisees when administering, scoring, and interpreting assessments. Specifically, professionals and supervisees, according to both the ETA and the PTA, should be qualified to administer, score, and interpret tests and assessments as demonstrated by their education, training, experience, and credentials.

Standard 10.15 from PTA states that interpretation of test results for diagnostic purposes should be based on multiple sources of data. Similarly, Standard 12.10 from ETA states that decisions which may have a major impact on students should take into consideration relevant information in addition to single test scores. Both standards suggest the importance of providing collateral information when using test scores to make decisions that

will have a major impact on an individual. In most cases, multiple data sources will often enhance the appropriateness of decision making.

Both ETA and PTA emphasize that test users should be familiar with the evidence of the validity and the reliability from the test when drawing inferences from test scores. Standard 10.2 requires the test users to know the validity and reliability of the test scores that are supported by logical analysis. Standard 12.14 from the ETA also states that professionals should be familiar with the reliability and validity of the test scores for their intended purpose as well as test fairness.

Regardless of whether tests and assessments are used in combination for multiple purposes or, when tests are packaged together for purposes of administration, evidence should be provided for each purpose of the test or for the package of the tests. Standard 10.16 states that, when tests are used in combination with one another, professionals should review the evidence for combining tests and Standard 12.2 states that one source of evidence for a test is not sufficient when the test is used for multiple purposes. It is to the task of the testing professionals to obtain evidence for every instance of the use of the assessment.

Finally, Standards 10.4 and 12.4 focus on the relationship of the test with the construct/knowledge the test is supposed to capture. Both standards state that evidence should be provided to support the extent to which the test samples the range of knowledge of the assessment and whether it meets its intended purposes. Also it is important to state explicitly the target domain that the test represents and those aspects that the test fails to represent (as in Standard 12.4).

2.2. Specific Differences between the Two Chapters

Despite the similarities that ETA and PTA share, there are also many differences between them. Chapter 10 (PTA) contains more emphasis on the test security than does Chapter 12 (ETA), based on their content. Test security is the sole topic of Standard 10.18 of PTA, which emphasizes that the obsolete versions of any test should not be kept available for the public. Standard 12.16 is the only part of Chapter 12 which discusses protocols for test security. The commentary note in Standard 10.18 of PTA provides extensive instruction on how to keep test information secure but no explicit standard and example are mentioned in ETA. However, although the ETA standards do not provide details about test security procedures, testing organization provides such information at the operational level. The North Carolina Department of Public Instruction (NCDPI, October 2019), for example, provides testing security protocols in extensive detail. This may indicate that test security level largely depends on the stakes of the test in educational settings whereas the stakes of assessment in psychological settings are usually very high.

ETA describes the general procedures of test design and

development (TDD) in educational assessment settings, however, PTA emphasizes selection of psychological tests for a particular test taker. ETA addresses topics such as minimizing the negative consequences of a test (Standard 12.1) as well as obtaining the evidence of validity and reliability for a test with multiple purposes (Standard 12.2). The aspects discussed in the TDD cluster of Standards from the ETA chapter also overlap with the general TDD chapter in Chapter 4 of the *Standards* (AERA et al., 2014). For example, the cluster provides general guidance for specifying a test (p. 83), developing and reviewing items (p. 85), and revising the test (p. 87). However, the standards of TDD specific to the ETA chapter are not mentioned in Chapter 4 given the nature and context of the educational field. On the contrary, no explicit cluster on TDD is available in Chapter 10 as Chapter 4 has already described the standards related to PTA.

PTA emphasizes test selection as well as concerns about topics related to choosing suitable tests for test takers (Standard 10.5) and choosing credible tests used for the diagnosis of different groups (Standard 10.6). There is no topic on test selection in the ETA chapter or in the related chapters under the “Part II: Operations” section in the *Standards* (AERA et al., 2014). This is not the case for the PTA chapter because professionals in the psychology field need to know the intended reasons to use a test in the assessment of a test taker (cf., Haladyna & Downing, 2004; Sackett, Borneman, & Connelly, 2008). They therefore need to choose not only relevant tests, but subsets of tests or diagnostic procedures to evaluate the test takers (p.152). The primary purposes of assessment in the educational setting are stated explicitly in the *Standards* (AERA et al., 2014): “[T]o know the students’ learning outcome, to understand the quality of the teaching they have received, and to make inferences about students’ learning level” (p. 184).

3. Discussion

In this section some of the issues and questions which may have occurred to the reader are explored. For the scope of this summary I chose to focus on two of the issues.

3.1. Comparisons of the Chapters between the 1999 and 2014 *Standards* (AERA, APA, & NCME, 1999, 2014)

Retrospective comparisons with the *Standards* published in 1999 indicate that the standards from the two chapters of the 2014 version are more organized in light of the categorization of standards into particular themes of testing practices (see Table 1). Previous standards without commentary notes have been extended, with the addition of elaborative comments regarding the stated standards, and several separate yet repetitive standards from 1999

have been combined into more general standards in the current version.

In the 2014 chapter on PTA, the purposes section now includes testing for social identity and psychological health in addition to the purposes of testing for personal awareness, growth, and action which existed in the 1999 *Standards* (AERA et al., 1999). The additional purposes of PTA are deemed important given the significant demographic changes globally (Lesthaeghe, 2014); see Teixeira, Frey, and Griffin (2015) for changes in North America; see also recently added examples in the comment under Standard 10.1 (AERA et al., 2014) and the highlighted issues in fairness in testing. Significantly, Standard 10.11 explicitly highlighted the importance of compliance with law enforcement in reporting test scores when required and exemplified the circumstances under which adherence to the 2014 *Standards* (AERA et al., 2014) should be enforced. This is not the case for the 1999 *Standards* (AERA et al., 1999) in which “compliance is less black and white in the 1999 *Standards*, and can be viewed as positive and negative by advocates and critics of testing.” (Camara & Lane, 2006, p.36).

As stated in the introductory chapter of the *Standards* (AERA et al., 2014), the ETA chapter was “rewritten to attend to the issues associated with uses of tests for educational accountability purposes” (p. 4). In some of the standards in the ETA chapter, repetitive words in the 1999 version were revised and specific terminologies (e.g., dates of administration, test takers’ age) were generalized (e.g., contextual information, demographics). One instance of a profound revision existing in the 2014 chapter concerns the issue of fairness in testing in ETA. The newly added Standard 12.3 and the revised Standard 12.13 call for targeted subgroups to have an equal chance to demonstrate proficiency on a test by considering all relevant steps of the testing process, such as the use of technology, accommodations, and modifications, which was not highlighted in the 1999 *Standards* (AERA et al., 1999). Also, the topics on reporting test scores have been significantly revised and polished, likely to attempt to achieve more accountable testing practices (Smith & Fey, 2000) (cf., Battistin, 2016).

3.2. Questions about and Issues concerning about the Two Chapters

As mentioned previously, the ETA chapter does not have a standard dedicated to test security as the PTA chapter does. With the increasing trend of computer-based assessments (Boevé et al., 2015; Boitshwarelo, Reedy, & Billany, 2017) and with the addition of such a substantive topic in the *Standards* (AERA et al., 2014; Camara & Lane, 2006), online test security may become a more pronounced issue (see also Lissitz & Jiao, 2012). This is an issue that should be addressed now (cf., International Test Commission (ITC) guidelines on Computer-Based

and Internet Delivered Testing (ITC, July 2005) and on the Security of Tests, Examinations, and Other Assessments (ITC, July 2014)) but should also be considered for future versions of the *Standards*.

Although both chapters present a comprehensive summary of the standards for their respective fields, there may exist standards in one field that should or could be explored in the other field. Worth noting is that standards that provide value not only to their intended area but can be generalized to others as well. For example, Standard 12.12 of ETA mentions that if two tests from a person are compared, the extent to which the two test statistics are very similar and the standard error of the test scores should be considered. This issue is not mentioned in Chapter 10 thus triggers another question concerning whether it should be addressed in PTA as professionals may use multiple tests to make a decision on an individual (Standard 10.6).

4. Conclusions

This paper’s cross-chapter comparisons of the 2014 *Standards* (AERA et al., 2014) aim to provide to test users and graduate students interested in testing and assessment some holistic insights regarding the focus, similarities, and differences of testing practices in the psychology and educational fields. With some of the issues and concerns raised, it is hoped that graduate students will be able to pursue their research and professional endeavors by adhering to the latest standards and guidelines. It is hoped that they would be able to generate significant research questions that can contribute to the field of testing and assessment in education and psychology. Future work may continue the comparison between the two chapters and Chapter 11: The Workplace Testing and Credentialing chapter, as discussion on certification/licensure and employment testing is often integrated with educational and psychological testing assessment, especially in the context of higher education (Sackett et al., 2008).

The brief comparison of the 1999 and 2014 versions may prove useful for test users and graduate students so that they will be aware of the progression of the *Standards* and are also cognizant of the latest testing standards, development, and agenda especially in the North American testing context. Although there remain improvements to the *Standards* (AERA et al., 2014) which may be made in the future, it is important for all novice and experienced testing practitioners, scholars, and users in the discussed fields to refer consistently to the standards and comply with most, if not all, of the provided guidelines to ensure sound and practical testing practices.

Acknowledgments

The author would like to sincerely thank Prof. Micheline Chalhoub-Deville, Juanita Hicks, and Yan Fu

from the Educational Research Methodology department, School of Education, at The University of North Carolina at Greensboro, North Carolina, for their feedback and suggestions for improvements.

REFERENCES

- [1] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1966). Standards for educational and psychological testing. American Psychological Association.
- [2] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1974). Standards for educational and psychological testing. American Psychological Association.
- [3] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). Standards for educational and psychological testing. American Psychological Association.
- [4] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). Standards for educational and psychological testing. American Educational Research Association.
- [5] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. American Educational Research Association.
- [6] American Educational Research Association, & National Council on Measurement in Education. (1955). Technical Recommendations for Achievement Tests.
- [7] American Psychological Association. (1954). Technical Recommendations for Psychological Tests and Diagnostic Techniques.
- [8] Battistin, E. (2016). How manipulating test scores affects school accountability and student achievement. *IZA World of Labor*, 295, 1-11. <https://doi.org/10.15185/izawol.295>.
- [9] Boevé, A. J., Meijer, R. R., Albers, C. J., Beetsma, Y., & Bosker, R. J. (2015). Introducing computer-based testing in high-stakes exams in higher education: Results of a field experiment. *PLoS ONE*, 10(12). <https://doi.org/10.1371/journal.pone.0143616>.
- [10] Boitshwarelo, B., Reedy, A. K., & Billany, T. (2017). Envisioning the use of online tests in assessing twenty-first century learning: A literature review. *Research and Practice in Technology Enhanced Learning*, 12(16), 1-16. <https://doi.org/10.1186/s41039-017-0055-7>.
- [11] Bornstein, R. F. (2017). Evidence-based psychological assessment. *Journal of Personality Assessment*, 99(4), 435-445. <https://doi.org/10.1080/00223891.2016.1236343>.
- [12] Camara, W., & Lane, S. (2006). A historical perspective and current views on the Standards for Educational and Psychological Testing. *Educational Measurement: Issues and Practice*, 25, 35-41. <https://doi.org/10.1111/j.1745-3992.2006.00066.x>.
- [13] Camara, W. J., Nathan, J. S., & Puente, A. E. (2000). Psychological test usage: Implications in professional psychology. *Professional Psychology: Research and Practice*, 31(2), 141-154. <https://doi.org/10.1037/0735-7028.31.2.141>.
- [14] Groth-Marnat, G., & Wright, A. J. (2016). *Handbook of psychological assessment* (6th ed.). John Wiley & Sons, Inc.
- [15] 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>.
- [16] International Test Commission. (July 2005). *International Test Commission Guidelines on Computer-Based & Internet Delivered Testing*.
- [17] International Test Commission. (July 2014). *International Test Commission Guidelines on the Security of Tests, Examinations, and Other Assessments*.
- [18] Lesthaeghe, R. (2014). The second demographic transition: A concise overview of its development. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 111(51), 18112-18115. <https://doi.org/10.1073/pnas.1420441111>.
- [19] Linn, R. L. (2006). The Standards for Educational and Psychological Testing: Guidance in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 27-38). Lawrence Erlbaum Associates, Inc.
- [20] Lissitz, R. W., & Jiao, H. (2012). *Computers and their impact on state assessments*. Information Age Publishing.
- [21] North Carolina Department of Public Instruction (October 2019) *Testing Security Protocol and Procedures for School Personnel*.
- [22] Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2008). High-stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychological Association*, 63(4), 215-227. <https://doi.org/10.1037/0003-066X.63.4.215>.
- [23] Smith, M. L., & Fey, P. (2000). Validity and accountability in high stakes testing. *Journal of Teacher Education*, 51(5), 334-344. <https://doi.org/10.1177%2F0022487100051005002>.
- [24] Teixeira, R., Frey, W. H., & Griffin, R. (2015). States of change: The demographic evolution of the American Electorate, 1974-2060. <https://cdn.americanprogress.org/wp-content/uploads/2015/02/24060014/SOC-reportAugust15.pdf>.