LEARNING PROGRESSIONS AND ONLINE FORMATIVE ASSESSMENT NATIONAL INITIATIVE

FINAL REPORT – ATTACHMENT 6

KEY ADVANCES IN EDUCATIONAL MEASUREMENT

acara AUSTRALIAN CURRICULUM, ASSESSMENT AND REPORTING AUTHORITY
Education Services Australia
aitsl Australian Institute for Teaching and School Leadership Limited

Contents

This paper was prepared by Dr Timothy O'Leary, Research Fellow, Melbourne Graduate School of Education and Director, Educational Data Talks at the request of the project team.

The views in the report are those of Dr O'Leary and do not represent the views or official positions of the project team.

1 Overview of report

The purpose of this report is to provide a user-friendly, relatively jargon free, overview of recent advances in the academic field of Educational Measurement.

1.1 Methodology

To gain an overview of advances in the field of Educational Measurement this report drew on three primary sources.

1. several international experts were contacted to provide their insights into developments in the field of Educational Measurement. In total five academics were generous in their responses and guidance. See Appendix 1 for their credentials.
2. article titles and abstracts from major journals related to Educational Measurement were reviewed from the last fifteen years to identify key themes. See Appendix 2 for a brief overview of the journals considered.
3. a brief literature review was conducted based upon expert feedback and initial read of the literature.

Insights from these sources were used to synthesis an overview of key developments in the field of Educational Measurement over the past fifteen years.

1.2 Structure of report

This report has been structured to include:

1. a summary of key advances in Educational Measurement
2. a summary of implications for current practice
3. a concluding comment
4. a full appendix outlining experts consulted, journals reviewed, and a reference list.

2 Key advances in Educational Measurement

Within the field of Educational and Psychological Measurement, there has been considerable movement in the last two decades. Much of the evolution that has occurred has been the result of a natural, technical progression associated with improvements in statistical methodology and technique. In addition, there have also been several advances that have taken the field of Educational Measurement significantly, and meaningfully forwards, in ways that are set to improve not only the field itself, but also the outcomes for which the field works (i.e. ultimately, improving learning outcomes for students).

This section presents a summary of the key changes and advances observed in the academic literature related to educational and psychological measurement and based upon expert input. These changes are captured under the themes of: Enhanced Test Development, Delivery and Scoring; Expanded Domains of Measurement; and, Validity and Score Reporting.

2.1 Enhanced test development, delivery and scoring

Over recent years, the nature of test development has changed significantly. For example, items for large-scale tests are increasingly created and assembled automatically by computer algorithms, as opposed to being manually developed. Such automated approaches to item construction have two key benefits. Firstly, automation can produce items in more cost-effective ways than traditional methods (Mosh, Simpson, Bickel, Kellogg and Sanford-Moore, 2019). Also, automation can produce items in sufficient volume to address potential security concerns (Gierl & Haladyna, 2012; Gierl & Lai, 2012; Luecht, 2012). In this context, automated test assembly (ATA) is an area which has

also made important progress in recent years (van der Linden & Diao, 2011). ATA involves an algorithmic approach to the selection of items to create parallel test forms from a large item banks and has become possible through methodological and technical advances. Several practical worked examples, using Microsoft Excel, can be found in Cor, Alves, and Gierl (2008, 2009). Automated item construction and automated test assembly are both developments that have the capacity to improve the cost effectiveness of assessment development. Moving forward, they should be considered a key feature of the development process for the Learning Progressions and Online Formative Assessment National Initiative (the initiative).

Evolving technology has also significantly impacted upon how assessments are conceived of and administered. For example, traditional paper-and pencil tests are quickly becoming redundant as assessments are now increasingly being designed as adaptive and delivered online, employing dynamic and interactive tasks and simulations (e.g. Gierl & Haladyna, 2012). Adaptive assessments have marked a move forward in Educational Measurement as they allow assessment to be designed in a manner that appropriately challenges all students which results in a more efficient and accurate testing process. Further, computer-based assessments, including adaptive tests, have been able to take advantage of innovations in item format, response action, media inclusion, level of interactivity, scoring, and communication of test results (Parshall, Spray, Kalohn, & Davey, 2002; Zenisky & Sierci, 2002). Indeed, advances in technology has seen many new item and task formats offered to capitalize on advances in technology and the learning sciences (Scalise and Gifford, 2006; Zenisky & Sierci, 2002). Multistage testing, also called computerized adaptive sequential testing (CAST) has also emerged as test administration methodology of interest. CAST is a process by which examinees are directed to later tests based upon their performance in an earlier test (for example, examinees who fare poorly on initial tests can complete easier levels in subsequent stages). Multistage tests are like computerised adaptive tests but have important differences (Mead, 2006) such as using groups of items, called testlets or panels, as opposed to individual items to create a test in stages (Luecht & Nungester, 1998).

Another key development has been the capacity to create and deliver virtual immersive performance assessments via rich simulations. Such assessments have several benefits as they alleviate the need for extensive training for administering tasks, the need for providing materials and kits for hands-on tasks (which has the potential to reduce equity issues resulting from lack of resources), and, potential safety issues (Midura-Clarke & Dede, 2010). In addition to simulations, game-based assessment has also emerged as an option for providing rich opportunities for assessing students. Recent work at Pearson, highlighted by Dicerbo (2010), has focused on aligning evidence from games to learning progressions that can be used to provide evidence of student mastery. This allows for games to be designed so that learners make meaningful choices aligned to the levels or stages of the progression. A promising feature of game-based assessments is that they can be designed with participant engagement and motivation in mind. This is important, as there is evidence suggesting that assessments in which students are more engaged and motivated result in more valid inference. Further, the types of scenarios used have the capacity to offer better evidence for constructs, particularly those that are difficult to measure in traditional ways. It should be noted though, that games are not necessarily an efficient means of assessment; in fact, asking a multiple-choice question can be much quicker. The most appropriate use of game-based assessment is, therefore, in formative situations where the game is both a learning and assessment tool.

The capacity to deliver large scale assessments with technology has been another significant shift within the field of Educational Measurement. This step has allowed the measurement of more traditional constructs (i.e. literacy and numeracy) to be undertaken more efficiently and effectively by accelerating assessment presentation, data collection and reporting (Bennett, 2018). There are several good examples of such assessments. For example, the Program for International Student Assessment (PISA) has been administered to approximately 400,000 fifteen-year-old students across 57 countries (Organization for Economic Cooperation and Development [OECD], 2017). ePIRLS, the digital component of the Program in International Reading Literacy Study, was given to 85,000 fourth grade students across 16 education systems. In the United States of America, the National Assessment of Educational Progress (NAEP) administered its reading and mathematics assessments on tablets to 150,000 students, and 20,000 in writing assessment. Notable to the NAEPs approach to implementing technology within their educational assessments has been their structured approach beginning with using research to drive small scale operational measures before progressing to larger assessments and samples (Bennett, 2018). Locally, the National Assessment Program (NAPLAN) has begun its transition to online assessment with the assessment being delivered to increasingly larger audiences in 2018 and 2019. It should be noted that whilst there are clear efficiency benefits to delivering large scale assessment

online there are also issues and/or concerns that have been raised. For example, as is well known in Australia, there have been concerns related to fairness, reliability, comparability, and ultimately validity of scoring, due to technology issues that have hindered delivery in 2019.

As test development and item construction has evolved, so too has item scoring. Automated score of written content is an area which has seen significant advances in recent years. For example, the Educational Testing Services (ETS) have conducted significant research on accurately scoring written content over the past two decades. Their work has moved from utilising natural language processing techniques, requiring significant human effort, through to machine learning techniques which needs far less human intervention. There are numerous examples of ETS's work in this area including: automated scoring of formative assessment of scientific argumentation (Mao, Liu, Roohr, Belur, Mulholand, Lee and Pallant, 2018); using neural networks for short answer scoring (Riordan, Horbach, Cahill, Zesch and Lee, 2017); and, automatically scoring tests of proficiency in music instruction (Madnani, Cahill & Riordan, 2016). A key sub-set of automated scoring that is also a major step forward for automated scoring of written responses is automatic short answer grading, which focuses on assessing short natural language responses to questions. This has been an important development as short answer questions have been recognized as a tool to perform a deeper assessment of the student's knowledge than, for example, multiple choice questions.

There are clear implications for the initiative that arise from developments in test development delivery and scoring. Firstly, with evolving options for assessment construction and delivery Australia should be looking to develop assessments that are adaptive and multi-staged in nature and utilising richer item format, response actions, media inclusions and interactivity. We should also explore if and how game-based and/or virtual performance assessments might be incorporated to facilitate richer assessment opportunities. Secondly, online delivery of assessments clearly creates potential for more efficient and effective delivery, analysis and reporting of assessments which has the capacity to accelerate improvement in student learning by providing actionable information to teachers in a timely manner. That said there are potential risks to validity that we must be sure to mitigate effectively. Ultimately, utilising automated approaches to the scoring of writing content, including short responses, is worth considering as it provides the capacity to perform deeper assessment of student knowledge.

2.2 Expanding measurement domains

Contemporary views on learning now highlight that deep learning occurs not simply by accumulating knowledge, but through using and applying knowledge as one engages in discipline-based activity (Harris, Krajcik, Pellegrino & DeBarger, 2019). Consequently, those concerned with education policy and practice have shifted their priorities toward supporting deeper learning by emphasizing the importance of students' ability to apply knowledge in subject areas. As a result, designers of student assessments have begun to follow suit and are taking up the challenge of creating a new generation of assessments with new types of tasks and situations that call upon students to demonstrate well integrated learning (Harris, Krajcik, Pellegrino & DeBarger, 2019).

Advances in technology have enhanced the potential to measure the core conventional skills of students. These enrichments focus on making proxy measures more closely linked with latent traits by increasing the amount of information that is captured. Technology can also help better understanding students' problem-solving processes by capturing task actions in addition to the usual test responses (Vista & Care, 2017). Technology driven assessments have also permitted the measurement of '*newer*' competencies (i.e. writing on computer, reading in hypertext environments, collaborating with remote partners in virtual spaces, and executing problem-solving processes) that are increasingly meaningful in terms of success in education, the workforces and meaningful citizenship that have not been possible to assess with traditional approaches to educational measurement (Bennett, 2018). Further, the integration of digital technologies within assessment systems has allowed for the use of game-based assessment as a methodology for providing assessment of multidimensional learner characteristics (cognitive, metacognitive and affective) using authentic digital tasks (e.g., games and simulations) (Shute, Leighton, Jang, Chu, 2016). The capacity to create and deliver virtual immersive virtual performance assessment, via rich simulations, has also emerged to complement existing standardised approach to assessment by assessing skills not typically captured by item-based assessment or hands-on real-world performance assessments.

In recent years, newer models of assessment have also emerged that provide greater insight into examinees. For example, cognitive diagnostic assessment (CDA) is a relatively new development in psycho-educational measurement which helps assessment researchers examine test takers' mastery of specific sub-skills with much greater specificity than more traditional models (i.e. Rasch or IRT models). This can be useful for providing much more fine-grained diagnostic information about test takers' degree of mastery of various defined sub-skills (Lee & Sawaki, 2009). Whilst CDA development was largely motivated by the need for new formative assessment methods, the technique has been retrofitted to norm-referenced tests (Jang, 2005, 2008) and responses from IRT based assessment forms in a move that potentially improves the actionability of the scores derived for examinee responses (Liu, Huggins-Manley & Bulut, 2017). It should be noted that whilst CDA is an emergent area of educational measurement in theory, applications beyond research are relatively sparse (*personal communications* Professor Derek Briggs).

There are two key implications for the initiative that arise from an expanding range of constructs being assessed. Firstly, Australia should look to build a tool to assess a comprehensive range of skills including traditional domains such as reading comprehension and mathematics augmented with emerging constructs. This should ensure that we are supporting our education systems to not only improve our students learning outcomes across the traditional fundamental skills but also enhance them in the emerging constructs of 21st century skills. Whilst cognitive diagnostic assessment is an emerging facet of educational measurement, particularly given its potential to yield rich diagnostic feedback, it is worth considering further research to determine its applicability to the initiative.

2.3 Validity, validation and score reporting

Validity of assessments has been a central concern for test development and test developers for the over 50 years. Since the late 1980s the focus of validity theory has shifted from being about the test towards the proposed interpretation and use of test scores (Messick, 1989). Since this shift in focus the work of Kane (1992, 2006, 2013) and Mislevy, Steinberg and Almond (2002) have contributed to improvement in the process of assessment design and validation. Mislevy et al's 2002 work on Evidence-Centered Design has provided a process to ensure a rigorous framework for developing assessments that measure their intended constructs and yield the evidence needed to support claims drawn from the results. Kane (1992, 2006, 2013) has provided a robust argument-based approach to validation which has been adopted in the validation process by many participants in the test development field. Bennett (2010) has also proposed theory of action as an effective methodology for subsuming validation efforts for assessment, in a manner that provides greater prominence to the effects of assessment over the purely technical aspects of test development.

As the focus of validity has more formally focused upon interpretation and use of test scores than score reports, the many and varied means by which scores are conveyed to their intended audience, have become an area of academic interest. This is because score reports are the main visible outcome of the complex process of testing and, as such, integral to the communication between test developers and their audience. As such, the effectiveness of a score report can influence the decisions and actions of its target audience. This is a critical component in achieving the intended outcomes of assessment programs. This might seem like common sense, but today there is evidence that score reports are often misinterpreted and misused; one only needs to look at how the Australian media reports upon and interprets NAPLAN to gain insight into how scoring and assessment can be distorted. Effective, intentional score reporting has the capacity to be the panacea to this problem. A good example of the epidemiology of modern score use, which advises a proactive stance on anticipating how scores will be used, is that of Ho (2013).

Now, thanks to academics including Zenisky and Hambleton (2012), Roduta Roberts and Gotch (2019), and, O'Leary, Hattie, and Griffin (2016) there are increasingly clear approaches to score report design and evaluation. Further, score reporting is increasingly seen as a significant consideration from the outset of test design and development, as opposed to a post development addendum. An excellent example of a score report design process embedded into the test development is that of e-asTTLE in NZ (see Brown, O'Leary & Hattie, 2019). The underlying message from this research is that score reporting needs to be a key aspect of the design process from the beginning stages of a testing initiative. In practice, this means that a clear set of tangible expected outcomes need to be mapped and subsequently unpacked into intended interpretation and use of scores from a testing initiative. This information can then be used to concurrently drive both score report design and assessment development. This is critical as it ensures that both the

underlying assessments (i.e. the items and tests) adequately support the intended interpretation and use of the scores generated and that the reports effectively communicate their scores and intended message to the audience.

Other aspects of score reporting, such as sub-score reporting, have also been prominent in the literature. A sub-score is a sub-domain or constituent part of a score. For example, in Australia students who complete NAPLAN receive an overall numeracy score but, in some states, a sub-score for the individual domains are reported to schools. Reporting sub-scores is an area of interested because it creates potential to enhance diagnostic value of tests that cover multiple underlying traits (Feinberg & Wainer 2014; Sinharay, 2010). One of the key questions regarding sub scores though has been concerns centred upon their reliability as they only have value if they are reliable enough to justify conclusions drawn from them and if they contain information about the examinee that is distinct from what is in the total test score (Feinberg & Wainer, 2014). Recent work by academics in this area have provided guidelines for reporting and interpreting subscores (Sinharay, 2010; Yao, 2010; Feinberg & Jurich, 2017).

There is one key implication for the initiative that arises from the evolution of validity theory, validation, and score reporting. Given that validity is a fundamental consideration for test development the initiative must ensure a robust methodology from the outset. A good start would be look to Bennett's (2010) advocation of using a theory of action approach as a means to integrate the technical aspects of validation with an evaluation of intended outcomes of the assessment program. Fundamental to this is prioritising score reporting as a component of the initiative test.

3 Summary of implications

The key implications for the Learning Progressions and Online Formative Assessment National Initiative (the initiative) that have been raised throughout the previous section are summarised below.

3.1 Implication 1 - Cost effective assessment development

To ensure cost effective assessment development automated item creation and automated test assembly should be considered as features of the development process for the initiative.

3.2 Implication 2 - Innovative test design

With evolving options for assessment construction and delivery, the initiative should aim for the tool be adaptive and multi-staged in nature, and richer in terms of the item format, response actions, media inclusions and interactivity. It is also worth considering if and how game-based and/or virtual performance assessments might be incorporated.

3.3 Implication 3 - Automated scoring

Utilising automated approaches to score of writing content, including short responses is worth consideration as it provides the capacity to perform deeper assessment of student knowledge in a more efficient manner.

3.4 Implication 4 - Diversified constructs

With a clear growth of constructs being assessed, the initiative should look to build capacity to assess as broad a range of constructs as possible including, but not limited to, traditional concepts such as reading comprehension and mathematics augmented with emerging constructs. Further, whilst cognitive diagnostic assessment is an emerging facet of Educational Measurement it is worth considering further research to determine its applicability to the initiative.

3.5 Implication 5 - Validity and score reporting

Given that validity is a fundamental consideration for test development initiative must ensure a robust methodology from the outset. Bennett's (2010) theory of action approach offers a means to integrate the technical aspects of validation with an evaluation of intended outcomes of an assessment program. Score reporting must be prioritised as a component of the design and development process.

4 Concluding thoughts

The Learning Progressions and Online Formative Assessment National Initiative (the initiative) is an ambitious gambit and should aim to be bold. Through its creation, Australia should look beyond simply building an assessment tool, towards building a world leading assessment platform that seeks to assess a diverse range of competencies.

Rather than simply transitioning traditional standardised assessment online the initiative should aim for the tool be adaptive and multi-staged in nature. The initiative should also look to integrate emerging options related to item formats, media inclusion and level of interactivity. Here, it is worth considering if and how game-based and/or virtual performance assessments might be incorporated. The initiative should also look to integrate methods of automation to support item development assessment construction and scoring.

Given the emergence of score reporting as a fundamental consideration in test development, how the initiative will report results will be of considerable importance to ensuring the intended outcomes of the initiative are fully met. This means that score reporting needs to be a key aspect of the design process from the outset of the initiative. Bennett's (2010) suggestion regarding the use of theory of action is an important consideration which should form the basis for developing the initiative. In particular, the initiative and its intended outcomes need to be mapped and subsequently unpacked into intended interpretations to be made and actions engaged with by the various stakeholder and intended audience for the assessment. This information can then be used to guide development of the initiative.

Finally, we should see the development of the initiative as more than simply developing a set of assessments to improve outcomes for our students. This is no doubt a noble cause, but by building a world leading platform that draws upon best practice in the field, we have a real opportunity to develop an international hub of excellence in educational assessment.

Appendix 1 – International experts consulted

Professor Jonathon Templin

Professor and E. F. Lindquist Chair in the Educational Measurement and Statistics Program
Department of Psychological and Quantitative Foundations
University of Iowa

Dr Jaqueline P. Leighton

School and Clinical Child Psychology
LEAFF Lab, Faculty of Education
University of Alberta

Professor Mark Wilson

Professor of Education
Graduate School of Education
University of California, Berkley

Professor Derek Briggs

Professor and Program Chair
Director of Research and Evaluation Methodology
School of Education
University of Colorado, Boulder

Laureate Professor John Hattie

Melbourne Graduate School of Education
The University of Melbourne

Appendix 2 – Journals reviewed

Journal of Educational Measurement

JEM (quarterly) publishes original measurement research and reports of applications of measurement in an educational context. Topics addressed in JEM reflect contributions to educational measurement that are related to the purposes of NCME. JEM is a vehicle for sharing improvements and innovations in educational measurement

Educational Measurement: Issues and Practice

The primary purpose of Educational Measurement: Issues and Practice (EM:IP) is to promote a better understanding of and reasoned debate on assessment, evaluation, testing, and related issues of practical importance to educators and the public. To this end, EM:IP publishes articles that are both timely and have the potential to advance the policies and practices of educational measurement.

Applied Measurement in Education

Because interaction between the domains of research and application is critical to the evaluation and improvement of new educational measurement practices, Applied Measurement in Education's prime objective is to improve communication between academicians and practitioners. To help bridge the gap between theory and practice, articles in this journal describe original research studies, innovative strategies for solving educational measurement problems, and integrative reviews of current approaches to contemporary measurement issues.

International Journal of Testing

The International Journal of Testing (IJT) is dedicated to the advancement of theory, research, and practice in the area of testing and assessment in psychology, education, counselling, organizational behaviour, human resource management, and related disciplines. IJT publishes original articles addressing theoretical issues, methodological approaches, and empirical research, as well as integrative and interdisciplinary reviews of testing-related topics and reports of current testing practices. All papers are peer-reviewed and are of interest to an international audience.

Applied Psychological Measurement

For more than thirty years, Applied Psychological Measurement has led the measurement field in presenting cutting-edge methodologies and related empirical research. Whether the setting is educational, organizational, industrial, social or clinical, Applied Psychological Measurement focuses on ways to use the most current techniques to address measurement problems in the behavioural and social sciences.

Educational and Psychological Measurement

Educational and Psychological Measurement publishes referred scholarly work from all academic disciplines interested in the study of measurement theory, problems, and issues. Theoretical articles will address new developments and techniques, and applied articles will deal strictly with innovation applications.

Appendix 3 – Reference List

Bennett, R., E. (2010). Cognitively Base Assessment of, for, and as Learning (CBAL): A Preliminary Theory of Action for Summative and Formative Assessment. *Measurement* 8, 70 – 91.

Bennett, R. E. (2018) Educational Assessment: What to Watch in a Rapidly Changing World, *Educational Measurement: issues and Practice*, 37(4), 7 – 15.

Brown, G. T. L, O'Leary, T. M., and Hattie, J. A. C. (2018) *Effective Reporting for Formative Assessment: The asTTle Case Example*. In D. Zapata-Rivera (ed) Score Reporting Research and Applications. New York, NY: Routledge

Cor, K., Alves, C., & Gierl, M. J. (2008). Conducting Automated test assembly using the premium solver platform version 7.0 with microsoft excel and the large-scale LP/QP solver engine Add-in. *Applied Psychological Measurement,* 32, 652–663.

Cor, K., Alves, C. & Gierl, M. (2009). Three Applications of Automated Test Assembly within a User-Friendly Modeling Environment. *Practice Assessment Research & Evaluation*, 14(14).

Devedzic, V., Tomic, B., Jovanovic, J., Kelly, M. Milikic, N., Dimitrijeivc, S., Djuric, D and Sevarac, Z. (2018) Metrics for students' soft skills. *Applied Measurement in Education*, 31(4), 282-296.

Dicerbo, K. (2010) Current state of game-based assessment. URL: **https://www.pearsoned.com/current-state-of-game-based-assessment/**

Embreston, S. E. (2016) Understanding Examinees' Responses to Items: Implications for Measurement, Educational Measurement: Issues and Practice, 35(3) 6 – 22.

Feinberg, R. A. and Jurich, D. P. (2017) Guidelines for Interpreting and Reporting Subcores, *Educational Measurement: Issues and Practice*.

Feinberg, R. A. and Wainer, H. (2014) When can we improve subscores by making the shorter? The Case Against Subscores with Overlapping Items. *Educational Measurement: Issues and Practice*. 33(3)

Galhardi, L. B & Brancher, J. D. (2018) Machine Learning Approach for Automatic Short Answer Grading: A Systematic Review. In G. R. Simari, Eduardo, F., Gutierrez-Segura, F., Rodriguez Melquiades, J. A. (Eds.) Advances in Artificial Intelligence - IBERAMIA 2018, 380-391.

Gierl, M. J., & Haladyna, T. M. (2012). Automatic item generation: An introduction. In M. J. Gierl & T. M. Haladyna (Eds.), Automatic item generation: Theory and practice (pp. 3–12). New York, NY: Routledge.

Gierl, M. J., & Lai, H. (2012). Using item models for automatic item generation. *International Journal of Testing*, 12, 273–298.

Harris, C. J., Krajcik, J. S., Pellegrino, J. W. and Debarger, A. H. (2019) Designing Knowledge-In-Use Assessments to Promote Deeper Learning. Educational Measurement: Issues and Practice, 338(20, 53 – 67

Ho, A. (2013). The Epidemiology of Modern Test Score Use: Anticipating Aggregation, Adjustment, and Equating. *Measurement*, 11, 64 – 67.

Jang, E. E. (2005). *A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL*. Unpublished doctoral dissertation, University of Illinois at Urbana Champaign.

Jang, E. E. (2008). A framework for cognitive diagnostic assessment. In C. A. Chapelle, Y.R. Chung, & J. Xu (Eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment (pp. 117-131).* Ames, IA: Iowa State University.

Jang, E. E. & Wagner, M. (2014). Diagnostic feedback in language classroom. In A. Kunnan (Ed.), Companion to language assessment (vol. 2, pp. 693–711). New York, NY: Wiley-Blackwell.

Kane, M. T. (1992). An argument-based approach to validation. *Psychological Bulletin*, 112, 527 - 535

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational Measurement* (4th ed.), pp. 17 – 64). Westport, CT: American Council on Education and Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), pp. 1 - 73

Kosh, A. E., Simpson, S. A., Bickel, L., Kellogg, M. and Sanford-Moore, E (2018) A Cost–Benefit Analysis of Automatic Item Generation. Educational Measurement: Issues and Practice, 38(1), 48 - 53

Lee, Y.W., & Sawaki, Y. (2009). Cognitive diagnostic approaches to language assessment: An overview. *Language Assessment Quarterly*, 6(3), 172-189.

Leighton, J. P., Chu, M-W., & Seitz, P. (2013). Cognitive diagnostic assessment and the learning errors and formative feedback (LEAFF) model. In R. Lissitz (Ed.), Informing the practice of teaching using formative and interim assessment: A systems approach (pp. 183–207). Charlotte, NC: Information Age.

Liu, R., Huggins-Manley, A. C. and Bulut, O. (2017) Retrolfitting diagnostic classification models to responses from IRT-based assessment forms. Educational and Pyshcological Measurement, 78(3), 357-383.

Luecht, R. M. (2012). An introduction to assessment engineering for automatic item generation. In M. J. Gierl & T. M. Haladyna (Eds.), Automatic item generation: Theory and practice (pp. 59–76). New York, NY: Routledge.

Luecht, R. M. & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement*, 35, 229-249

Madnani, N, Cahill, A. & Riordan, B. (2016) Automatically Scoring Tests of Proficiency in Music Instruction, *Paper in Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications,* 217–222

Mao, L., Liu, O. L., Roohr, K., Belur, V., Mulholand, M, Lee, H-S and Pallant, A (2018) Validation of Automated Scoring for a Formative Assessment That Employs Scientific Argumentation, *Journal of Educational Assessment,* 23 (2), 121–138.

Mead, A. D. (2006) An Introduction to Multistage Testing. *Applied Measurement in Education*. 19(3)18-187

Messick, S. (1989b) Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (2002). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3-67.

Mosh, A. E., Simpson, M. A., Bickel, L., Kellogg, M, & Sanford-Moore, E. (2019). A Cost-Benefit Analysis of Automatic Item Generation. Educational Measurement: Issues and Practice, 38(1), 48-53.

O'Leary, T., Hattie, J., & Griffin, P. (2016) *Design Principles for Action and Outcome Focused Score Report Design*, Presentation to 10th Conference of the International Test Commission, Vancouver, July, 2016

Organization for Economic Cooperation and Development (OECD). (2017). PISA technical report. Paris, France: Author. URL: http://www.oecd.org/pisa/sitedocument/PISA-2015-technical-report -final.pdf

Clarke-Midura, J. & Dede, C. (2010) Assessment, Technology, and Change. Journal of Research on Technology in Education, 42(3), 309-328.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). Practical considerations in computer-based testing. New York, NY: Springer-Verlag.

Riordan, B., Horbach, A., Cahill, A., Zesch, T. & Lee, C. M. (2017) Investigating Neural Architectures for Short Answer Scoring, *Paper in Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, 159–168.

Roduta Roberts, M., & Gotch, C. M. (2019). Development and examination of a rating scale to assess score report quality. *Frontiers in Education: Assessment, Testing, and Applied Measurement*. 4(20), 1-10.

Scalise, K., & Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing "intermediate constraint" questions and tasks for technology platforms. The Journal of Technology, Learning, and Assessment, 4(6).

Shute, V. J., Leighton, J. P., Jang, E. E. and Chu, M-W (2016) Advances in the Science of Assessment, *Educational Assessment*, 21(1), 34-59.

Sinharay, S. (2010) How often do subscores have added value? Results from Operational and Simulated Data. *Journal of Educational Measurement*, 47(2) 150 – 174.

van der Linden, W. J. and Diao, Qi (2011) Automated Test-Form Generation. Journal of *Educational Measurement*, 48(2), 206 – 222.

Vista, A. and Care, E. (2017) Education assessment in the 21st century: New technologies, Brookings Institute. URL: https://www.brookings.edu/blog/education-plus-development/2017/02/27/education-assessment-in-the-21st-century-new-technologies/

Yao, L. (2010) Reporting Valid and Reliable overall Scores and Domain Scores, *Journal of Educational Measurement*, 47(3), 339 – 360.

Zenisky, A. L. & Sierci, S. G. (2002) Technological Innovations in Large-Scale Assessment, *Applied Measurement in Education*, 15(4), 337-362.

Zenisky, A. L. and Hambleton, R. K. (2012) Designing Test Score Reports that Work: The Process and Practices for Effective Communication. *Educational Measurement: Issues and Practice*, 31(2) 1 – 48.