A Review of the Validity Evidence in Support of the Information and Communications Technology Literacy Assessment (ICTLA)

A Validity Audit Report Prepared for the World Bank's READ RAS Program Initiative

Howard T. Everson[1]

Center for Advanced Study in Education
City University of New York

January, 2013

[1] This report was prepared under an independent contractor agreement between the author and the World Bank. The opinions expressed here are solely those of the author and do not reflect the views of the Center for Advanced Study in Education or the City University of New York.

INTRODUCTION

Beginning in 2005 the World Bank, in collaboration with the READ RAS Program and the National Training Foundation, announced plans to develop a large-scale information and communications technology literacy assessment (hereafter the ICTLA) in response to a request from the Russian Federation's Ministry of Finance. Between 2005 and 2011 these groups have worked together to design and develop the ICTLA, a test designed to assess 9th grade students' proficiency in digital information management and other related technology skills required for success in the 21st century. The ICTLA scores would be used for "low-to medium-stakes decisions, both at the individual student level and aggregate level" (the later helping to inform educational policy and practice in this increasingly important area). The ICTLA sponsors have identified four possible uses for this newly designed assessment:

- *informing policy* through the aggregation of scores across subgroups of students, and allocating resources for supporting instruction in key ICT literacy areas;
- *guiding instruction* by using ICTLA scores to identify students' strengths and weaknesses and designing evidence-based instructional interventions;
- *identifying students* who could benefit from instruction in ICT literacy; and
- *evaluating* the efficacy of ICT literacy instructional programs.

The READ RAS Program requested a validity audit be conducted to review the analytic and empirical evidence gathered to date in support of the intended uses of the ICTLA. As background for the validity audit, a set of background documents were provided by the READ RAS Program which included summaries of empirical findings from an initial small field test of the ICTLA. Collectively, these documents describe the purpose and rationale for the ICTLA, and offer a high-level overview of the ICTLA. Supplementary documents include the *ICT Literacy Assessment Technical Manual*, and an accompanying report, the *ICT Literacy Assessment: Progress on Validity Evidence,* which together summarize the psychometric evidence developed to date for the ICTLA. These documents and the data and empirical results contained therein served as the basis for this validity audit.

Given the array of possible interpretations and uses of the ICTLA test scores outlined in the READ RAS test development plan, the challenge for the test developers and sponsors is to generate the evidence and marshal the arguments in support of the validity of the ICTLA. With eyes squarely on this challenge, this audit report has a threefold focus: (1) to outline a validity framework for interpreting the evidence gathered to date; (2) to review the existing empirical evidence and psychometric analyses—thereby evaluating the strength of the empirical evidence currently in-hand; and (3) to recommend steps for gathering additional evidence and strengthening the argument in support of the validity of the ICTLA. Our goal is to provide a validity audit report that informs subsequent discussions by the various ICTLA collaborators and stakeholders—including test developers, researchers, and others in the READ RAS Program's growing consortium of policymakers, educators, students and parents—about the potential utility of the ICTLA. With this in mind, we organized the audit report along the following lines. The very next section sets the stage for identifying and developing the various forms of evidence needed to support the validity of the ICTLA, and does so by describing contemporary perspectives on the validity of educational measurements. Our focus then moves to a more nuanced discussion, in section three, of the relevant standards, constructs, measurement perspective, and test-based claims identified in the ICTLA's test framework. Here we review the mix of arguments, empirical and psychometric evidence, and warrants, offered in support the validity of the ICTLA. Once we have a clear picture of the extent and quality of ICTLA's existing validity evidence, we then summarize this analytic and empirical evidence, and close by offering a series of recommendations for developing additional supporting documentation and ongoing validity evidence.

CONTEMPORARY PERSPECTIVES ON VALIDITY

The joint AERA/APA/NCME test standards (1999) and the *ETS Standards for Quality and Fairness* (ETS, 2002) referenced in the READ RAS Program documents frame the test validity argument largely in terms of construct validity (i.e., "the concept or characteristic that a test is designed to measure" (Messick,1989, p.5). That is, contemporary test standards reflect a construct centered approach to test validity (Kane, 2006; Messick, 1989; and Mislevy, 2009). This perspective on test validity draws heavily on the idea that the theoretical construct underlying the test design of the ICTLA is well represented by the observable test scores.

Following Messick (1994), it is helpful to view the validity of the ICTLA as “an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores” (Messick, 1989, p. 13). As others have noted more recently (Kane, 2006; and Mislevy, 2009), the challenge of a construct-centered approach to test design is the frequent lack of consensus—either theoretically or operationally--on the definition and operationalization of a test’s target construct.

Important work, however, has been done recently to refine and advance views of validity in educational measurement (see, for example, Haertel and Lorie 2004; Kane 1992, 2001, 2002, 2006; Mislevy, Steinberg, and Almond, 2003). Moving beyond a construct-centered view of validity, an even more contemporary, post-modern perspective calls for an interpretive validity argument. One that “specifies the proposed interpretations and uses of test results by laying out the network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the performances” (Kane, 2006, p. 23). In the case of the ICTLA, an interpretive validity argument ought to be developed to provide the backing for the assumptions set forth in the ICTLA framework documents. Kane (2006) and others (Haertel & Lorie, 2004; Mislevy et. al., 2003), distinguish, for example, between the interpretive argument (i.e., the propositions that underpin test score interpretation) and the collection of supporting evidence and arguments to back the propositions or claims that comprise the narrative of the interpretive argument. This view was elaborated by Mislevy, et al. (2003), who asserted that supportive validity evidence ought best be organized as a structured argument, a reasoned approach akin to a formal legal argument (Toulmin, 2003). Further, Mislevy suggests a sound validity argument provides strong evidentiary support for a particular claim or assertion, permitting it to withstand criticism.

This distinction is useful for framing the audit of the ICTLA’s validity evidence. Thus we adopted this approach to validity largely because it provides a practical application and direction for deconstructing the ICTLA validity framework. We begin by explicating the ICTLA claims of interest and then looking to the documentation and evidence for justification of those claims, all the time identifying the evidence (data, study results, and/or expert opinions) to support those claims. We also look, along the way, for warrants or justifications that connect the supporting evidence to the central claims and intended uses of the ICTLA scores. To be clear, the strategy is to locate and interpret the various forms of evidence and explain how a particular aspect or strand of the evidence supports a particular ICTLA claim of interest. Using this approach, we examined the argument structures supporting the ICTLA assumptions. In doing so we adapted Kane's (2006) framework by highlighting the following evidence strands and offering links to the overarching ICTLA validity argument. Hence, we looked at the following aspects:

Scoring inferences—review item and task scoring rules assigning scores to examinees′ test performances. For us scoring inferences rest on two assumptions: (i) the scoring rules are appropriate and reasonable; and (ii) these rules are applied correctly and accurately to students′ responses in all instances and administrations of the ICTLA.

- *Generalization inferences*—here we link interpretations of observed test scores to claims about a universe of observations represented by the ICTLA's items and tasks. These inferences also rely on two assumptions: (i) the sample of observations is, indeed, representative of the larger universe or target domain, for the ICTLA this is the educational technology standards (ISTE, 1998); and (ii) the field test sample of examinees and items is sufficiently large to minimize sampling error.
- *Extrapolation inferences*—this step examines the link between interpretations based on students' test performances to claims about their knowledge, skills and abilities in the information and communications technology arena. This interpretive argument also rests on two assumptions: (i) the test items and tasks draw on the competencies and proficiencies outlined by experts in the field; and (ii) there are no irrelevant sources of score variance that would otherwise bias score interpretations. And finally,
- *Decision inferences*—this is the use of the test scores to assign (or place) students in instructional programs or courses based on those scores. These inferences depend on at least two value-laden assumptions: (i) students; test performance relies strongly on the knowledge, skills and abilities developed in earlier courses and instructional opportunities; (ii) and students with underdeveloped competencies have a low probability of success on the ICTLA and are not expected to succeed in subsequent information and technology courses.

To restate the obvious, this post-modern approach to validity, one based on the development of an interpretive argument--a narrative weaving together claims, and supported by a collection of evidence linked to the claims by explicit warrants—was adopted to guide the validity audit of the ICTLA. Our audit rests on the core idea of educational test validity as an interpretive argument (Kane, 2006), a series of claims, each of which is backed by empirical evidence and connected by a narrative that explains why and how the evidence supports those claims. (Booth, Colomb & Williams, 2008).

THE ICTLA's CURRENT VALIDITY EVIDENCE

The discussion in this section of the paper represents the core of the validity audit. We begin by identifying the central claims of the ICTLA developers, and we then take a close look at the evidence needed to support each of the claims.

Claim 1: The Content of the ICTLA Aligns with the Relevant ICT Literacy Standards. The key issue here is whether the test designers and developers produced an assessment that, from a content standpoint, reflects the widely adopted ICT literacy standards promulgated by the developers of those standards, i.e., the *American Association of School Librarians* (AASL) and the *International Society for Technology in Education* (ISTE).

Evidentiary Support. Support for this key claim should come from test design documents describing how the evidence centered design task models were created and used in the test development process. Additional support may also come in the form of a series of studies employing accepted standards alignment methodologies found in the educational measurement literature (see, for example, Porter & Smithson, 2001; Webb, 1997), methods that permit an analysis of the test items and tasks against the standards themselves. We expect the relevant evidence would contain a mix of things, including use of assessment engineering principles (Luecht, 2013), evidence-centered design task models (Mislevy & Risconscente, 2006), test frameworks and blueprints, as well as a description of other standards alignment approaches that may have been used (Porter & Smithson, 2001; Webb, 1997). The intentional and appropriate use of professionally accepted design principles, practices and methodologies (for example, the use of an evidence-centered design for creating ICTLA items and tasks) would indeed support inferences from the tests' content and level of cognitive demand to the adequate representation of the ICT literacy standards.

Claim 2: The ICTLA Meets the Professional Standards of Validity, Reliability and Fairness. This fundamental claim is that the ICTLA meets the *Standards for Educational and Psychological Testing* (the joint standards) sponsored jointly by American Educational Research Association (AERA), the American Psychological Association (APA) and the National Council on Measurement in Education (NCME), as well as the *ETS Standards for Quality and Fairness*. This, in essence, is a claim about the quality assurance steps that were taken as the ICTLA was designed, created and subjected to pilot testing. The focal point is whether professional testing standards (i.e., test development principles and practices) were adhered to by the ICTLA development team.

Evidentiary Support. The evidence to support this claim is often developed through a systematic review of the technical manuals, framework documents, scientific reports and other psychometric documentation provided by the READ RAS Program Thus, documents and evidence supplied by the READ RAS Program will be examined in light of these prevailing professional standards which emphasize providing test users and other stakeholders with quality assurances of the validity, reliability and fairness of the ICTLA.

Claim 3: ICTLA Criterion Scores Indicate Proficiency. The ICTLA examinations offered to the students during the 9th grade, presumably, will have cut-scores set that will permit an inference that students achieving these scores are proficient with respect to their ICT literacy skills. Thus, this key claim, which includes using the ICTLA scores to classify students, guide instruction, and inform policy, is essentially about criterion-referenced inferences, and presumes the proficiency criteria are well defined and accurately mapped to the ICTLA score scale.

Evidentiary Support. In many ways this is the most salient and possibly most controversial claim of the entire ICTLA initiative. Consequently, it requires gathering substantial statistical, psychometric, and judgmental evidence that, cumulatively, allow for specifying the performance levels that reflect accurately students' ITC literacy levels. As we noted earlier, we have adopted Kane's (2006) validity framework which looks at a number of aspects of the test development process, including scoring methods and rules, interpretative links from the test items and tasks to the test scores as representation of the ICT literacy domain, links between ICTLA test scores to inferences about students' proficiency in the target domain, and the weight of the evidence to support instructional and placement decisions by educators. In addition, a variety of methods are available in the educational measurement literature—some empirical or statistical, others more judgmental—for setting cut-scores (Cizek & Bunch, 2007; Zieky, Perie, & Livingston, 2008), and the development team's reliance on these methods will be examined as well.

<u>Review of the Available Validity Evidence</u>

Below we review the nature and quality of the evidence provided by the READ RAS Program and its collaborators in support of the various claims made on behalf of the ICTLA. The collection of available evidence—documentary, analytical, and empirical—is discussed using the lens of the contemporary validity framework described earlier. Where appropriate we reference the issues and evidentiary sources outlined in the audit review template supplied by colleagues at the READ RAS Program (see Appendix A). And throughout, we attempt to draw connections to the generally accepted professional test development standards as they relate to validity, reliability, setting cut-scores, scaling and equating, and test score uses.

As previously noted, the ICTLA developers set out a number of *claims of interest* that we expect to support by reference to the evidence developed through analytic and psychometric studies: How well does the test content align with the accepted academic standards in the field? Whether professional test developments principles and practices were adhered to throughout the ICTLA's design and development phases? And, to what extent are these analyses, and the psychometric and statistical evidence, sufficient to support the score-based inferences proposed by the READ RAS Program?

Next, we take each claim, in turn, and describe and evaluate the evidence available in support of it.

Claim 1: The Content of the ICTLA Aligns with the Relevant ICT Literacy Standards. As we said earlier the issue here is whether the content of the ICTLA is aligned with the generally accepted ICT literacy standards. According to the documents supplied by the READ RAS Program, the ICTLA "focuses on cognitive problem solving, critical reasoning, and [the] critical reading skills associated with using elementary technology to handle information (see Zelman, et al., 2011, p. 3). Repeatedly throughout the documents the test developers state the ICTLA is intended to assess the seven cognitive processes of ICT literacy identified by the Educational Testing Service in 2007, and derived from the standards published by both the AASL and the ISTE in 2007. These standards—both from a content and cognitive perspective—are central to the test's framework according to the ICTLA's documentation.

Evidentiary Support. Given the centrality of the content and process standards, a key piece of evidence from a validity perspective would be documentation that makes clear and explicit the ICT literacy assessment framework—a guide document aimed at the test designers and developers that helps identify the cognitive process and content targets of the assessment. Examples of assessment frameworks are widely available in the literature (see, for example, NAGB, 2014) and have proven useful for understanding the connections between standards and assessment design. Unfortunately, a test framework was not among the documents and the materials we reviewed. More clearly, we were looking for something of a map or a test blueprint which would link ICTLA test items and tasks to the AASL or ISTE cognitive processes and content standards. On a somewhat related note, in a number of the source documents we found references to the intention to use evidence-centered design (ECD) methods in the early stages of the test development process. However, only one example of an actual ECD task model was found. The test design documents did not refer to. or explain, in a detailed, descriptive way how ECD processes were used to create the collection of ICTLA items and tasks, or how they mapped to the underlying ICT standards. We found no discussion of the linkages from the items to the task models, and from the tasks to the standards-based claims in the Program's audit documentation. Thus, it strikes us that substantially more validity evidence (i.e., a test framework document, examples of task models, items and tasks, as well as maps of items and tasks to standards) is needed to support the claim that the test is a valid representation of the ICT literacy standards.

Claim 2: The ICTLA Meets the Professional Standards of Validity, Reliability and Fairness. This fundamental claim establishes the bona fides of the ICTLA, i.e., that the design and development of the assessment meets the widely accepted standards in the testing profession.

Evidentiary Support. Typically, the evidence to support claims of this type is found in a well- developed and comprehensive technical manual, a document describing in detail how the test(s) was developed, the nature of the test's content and other statistical specifications, item construction methods, scoring procedures, as well as summaries of supporting psychometric studies. The prototypical technical manual captures and describes, in great detail, all the steps taken to develop the assessment. These steps, and their relation to the professional standards, often include: (1) identifying the purposes of the test; (2) delineation of the construct to be measured; (3) development of test specifications; (4) item and task development; (5) details of the field test studies undertaken; (6) description of the test administration procedures; (7) item analyses; (8) test assembly and evaluation methods; and (9) the development of score reports and other interpretative materials.

Although there was, in fact, some partial, often incomplete, information about the early field test that was conducted using students from Tartarstan and Thailand, the documents submitted for review by the READ RAS Program, in general, contained only fragmentary information directly relating how the professional test development principles and practices were used in the creation of the ICTLA. It could be that the ICTLA has been built on frameworks and technical specifications established by others doing earlier work in this field (the Educational Testing Service in the USA, for example). If these earlier efforts are (or were) part of the development process, the technical details of this early development effort ought to be reported as part of the technical documentation for this version of the ICTLA. In our view the ICTLA validity argument would be strengthened substantially by a carefully written and well-documented ICTLA technical manual.

Claim 3: ICTLA Criterion Scores Indicate Proficiency. The developers of the ICTLA examinations, presumably, have gone through a standard-setting process and arrived at cut-scores on the ICTLA scale that support the inference that students achieving these scores are proficient (or not) with respect to ICT literacy standards represented by the assessment. This work is referred to, albeit obliquely, by Zelman, et al. (2012) in their Table 2 which offers a detailed description of the five proficiency levels that were used to transform ICTLA scale scores into proficiency (or performance) level descriptors.

Evidentiary Support. Clearly, student proficiency claims, and the decisions based on those proficiency scores, are likely to be the most controversial claims of the entire ICTLA initiative. As such, the validity evidence to support these claims is considerable, often requiring the collection of large, rich examinee-level data, and multiple forms of analyses (including psychometric and statistical), and a well-documented standard-setting method. Keeping in mind Kane’s (2006) framework (i.e., appropriate scoring rules, use of data from large representative samples of students, and clear connections among and between the items, tasks, and process and content standards), the data from the ICTLA field trials, by necessity, bear a great deal of the evidentiary weight. In an evidence-based approach to test development and validation much rests on the quality and extensiveness of the empirical evidence gathered during the field trial phases. It is critical, therefore, that the field trials be well-designed and well-executed. Meeting this stiff design challenge, given the relatively small and unrepresentative samples used in the calibration study reported in document subtitled *Progress on Validity Evidence* and elsewhere (see Zelman, et al. 2012), will require, in our view, mounting another field trial data—one that gathers item-level data and other relevant background measurs from a larger, more representative samples of students.

With respect to the question of how proficiency levels were set, the documentation was again incomplete. The standard-setting methods used to arrive at these proficiency statements were not described in the collection of supporting documents provided by the READ RAS Program. Although a variety of applicable standard-setting methods are available in the literature—some empirical or statistical, others more judgmental (Beimers, Way, McClarty & Miles, 2012; Cizek & Bunch, 2007; Zieky, Perie, & Livingston, 2008)—none were pointed to as having been used to support of the validity for the claim that ICT literacy levels were arrived at systematically and through an evidence-based approach. If expert panels were employed, who were the experts and how were they chosen and trained for the standard-setting tasks? In the end, a good deal more descriptive information and documentation of the psychometric properties of the items/tasks is needed, along with a complete description of how the performance level indicators and descriptors were achieved, is needed to support the ICLA validity argument.

In the next section we offer a bit more detail and describe the limitations to the current field trial study, and discuss the strengths and weaknesses of the empirical evidence in light of those design constraints.

Apparent Limitations of the Calibration Study

By piecing together statements from three separate documents (i.e., the research paper by Zelman, et al. 2012; the *Progress on Validity Evidence* document; and a briefing paper reviewing preliminary convergent and divergent validity evidence), our overall impression is the field trial (or calibration study) relies on much too small a sample of students (n=395 students yielding only 277 useable cases), and therefore provides limited and inadequate empirical evidence to support the score- and test-based claims for the ICTLA. The samples—one using students from Tartarstan and the other students from Thailand—are not described completely enough to warrant strong generalizations to other populations of students and examinees. How were the students recruited? What do they look like in terms of gender, country of origin, academic achievement, etc.? Some of this information is presented in different places in the documents, but the technical manual and other documentation about the field trial (or calibration study) lacks clear and concise descriptions of how the samples were drawn, how the ICTLA was administered to the sample students and whether, for example, alternate forms of the test were developed and administered to equivalent groups of students in these two samples. As a consequence, the available data tell us little about the statistical characteristics at the item or task level. Did all of the items survive the field trials? Were some items and/or tasks too difficult? Are they all measuring well the underlying latent constructs?

Much more information is needed in terms of how many students took each item, how much missing data there was and how it was handled in subsequent analyses, and how item difficulty and discrimination parameters were estimated given the limitations of such small samples. As a consequence of these study design limitations, it is not prudent to say anything definitive about the item (or task) characteristics, the overall difficulty of the test, and the appropriateness of the proficiency levels set on the score-scale, or to make claims (draw inferences) about the actual ICT literacy skill levels of the larger universe of 9th grade students in those jurisdictions (i.e., school districts, countries or regions) targeted by the READ RAS Program. To further underscore the problem, as we mentioned earlier there appear to be relatively large amounts of missing data—not only with respect to the ICTLA scores but also to the other measures of individual differences that were administered. It is well known that problems of missing data, particularly when the *missingness* is proportionately large as is the case here, exacerbate item calibration estimates—biasing the item parameter estimates and contributing to the poor fit of item response theoretic models.

Nevertheless, it is the case that some relevant, and potentially useful, preliminary analyses were conducted in support of the ICTLA. Inter-item correlations, for example, were computed, and zero-order correlations of scores from self-report measures with ICTLA total scores were also reported. In addition, scale reliability indices were reported for the self-report measures, and preliminary factor analytic evidence was developed and reported to support the internal structure of the ICTLA. For the most part, these analyses and statistical indices are scattered among three documents—the Zelman, et al. 2012 paper, the *Progress on Validity Evidence* document, and the ICTLA Technical Manual. Perhaps more important, however, is that these analyses were conducted using very small samples (and for the most part the sample sizes for each of the analyses were not reported). As we noted earlier, the concern here is that the small sample sizes, coupled with the unclear—but evident—proportions of missing data at the item and student level, lead to item and test characteristics that are at best rough estimates, and at worse incorrect. The preliminary statistical and psychometric findings, and the samples on which they are based, need to be re-reviewed and integrated into a single document (preferably a detailed technical manual) that presents a more coherent picture of these preliminary results and pieces the evidence together, one with the other, to create a more complete validity argument in support of the ICTLA.

CONCLUSION

Constructing a validity argument to support the intended uses of the ICTLA requires the synthesis of evidence from a variety of sources. Some comes from carefully designed statistical and psychometric studies (e.g., statistical studies investigating the relationships among and between the ICTLA scale scores and other measures of individual academic performance and attitudes toward technology). Still other evidence comes from psychometric benchmarking studies that allow for predictions or projections from the test scales, and other evidence from investigations into the requisite set of academic knowledge and skills needed for success in courses (basic or advanced) that are designed to further students' ICT literacy skills. The collection of evidence in support of the validity of the ICTLA, a bold and innovative assessment to be sure, will require a substantial, ongoing commitment to research and development. A brief, and admittedly incomplete, set of recommendations for creating a validity studies agenda is offered below.

Recommendations

- *Develop an ICTLA Framework Document*. The ICTLA design team claims their assessment rests squarely on, and is derived directly from, the AALS (2007) and ISTE (2007) information and communication technology literacy standards. An ICTLA framework could be crafted to show how these content and cognitive process standards are reflected in the items and tasks that comprise any given form of the ICTLA. This framework would also allow from a clear explication (and linkage) of the evidence-centered design methods used and how those methods lead to the production of items and tasks that, individually and collectively, provide evidence of students' knowledge, skills, and abilities in this particular domain. The educational measurement literature, particularly with respect to large-scale assessments like NAEP, TIMSS and PISA, provides a number of examples of assessment frameworks and how they can and do support test validity (see, for example, the *2014 Technology and Engineering Literacy Framework for NAEP* published by the National Assessment Governing Board).

- *Document Test Development Steps*. Keeping in mind that the ICTLA is intended for widespread, large-scale use in the countries and jurisdictions within the umbrella of the READ RAS Program, the test development process is likely to be ongoing—year after year—as the initiative expands its reach. This will mean, inevitably, larger and larger item and task pools will have to be created, and multiple (parallel) test forms will be needed. The challenges of moving to large-scale assessment development are many, and include, for example, ongoing attention to item development, a strategy for field testing new items, the design of multiple test forms, and the need for equating and scaling studies to ensure fair and reliable tests are in use year after year. Careful documentation of the test development steps ensures the production of high quality test forms now and in the future. These documents can be used in the future to provide test users and other stakeholders assurances that professional test development standards were used throughout the development and implementation of the ICTLA.
- *Implement a Well-Designed Field Test.* As we said earlier, much of the validity evidence required to support the ICTLA ought to come from a carefully designed and well-executed field trial—one that includes large samples of students, and that exposes as many test items and tasks as possible to as many students as possible. Specifying, in detail, the design of a complex field test is beyond the scope of this audit report. However, colleagues at the READ RAS Program, in conjunction with outside experts, ought to conduct a debriefing that focuses on better understanding the limited resources and constraints that lead to the less than sufficient initial field test or calibration study. Once these issues are surfaced, we are quite confident a more robust field trial can be designed and carried out successfully.
- *Convene a Formal Standard-Setting Panel.* The ICTLA materials reviewed for this validity audit made clear that the sponsors of the assessment intend to report proficiency levels based on students’ performance along the ICTLA reporting scale. To assure test

users of the validity and utility of these proficiency levels, it would be appropriate to gather together content experts and other stakeholders for the sole purpose of the providing judgments about the appropriate number of proficiency levels required, and to identify the cut-scores that reflect these proficiency levels. The results of this standard-setting effort ought to be documented fully and included as part of the evidence to support the validity of the ICTLA scale scores and proficiency levels.

- *Publish a Detailed ICTLA Technical Manual.* The joint testing standards published by AERA, APA, and NCME call for test developers and assessment sponsors to compile the test development steps, field trial results and reports of other related statistical and psychometric studies. Typically these various forms of descriptive and technical documentation is written and crafted in the form of a technical manual. An outline (or preliminary draft) of an ICTLA technical manual was among the documents reviewed for this validity audit. That document ought to be revisited and redrafted to include all the descriptive and technical studies and reports available. Because validity is viewed as an ongoing process, many contemporary test developers and sponsors have turned to online publishing of their technical manuals so they can more easily incorporate new data as it becomes available.

Given the proposed uses of the ICTLA scores the immediate technical challenge is to outline in sufficient detail a validity studies agenda that supports the core assumptions underlying the READ RAS Program with respect to benchmarking ICT literacy skills. This complex, interrelated and multi-faceted set of assumptions presents the READ RAS Program with a formidable technical challenge as it moves ahead with the development and implementation of the ICTLA.

,

REFERENCES

AERA, APA, & NCME (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Association of School Librarians (2007). *AASL Standards for the 21st Century Learner*. Downloaded from http://www.ala.org/aasl/standards.

Booth, W.C., Colomb, G.G., & Williams, J.M. (2008). *The Craft of Research* (3rd. ed.). Chicago, IL: University of Chicago Press.

Cizek, G. J. & Bunch, M.B. (2007). *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests.* Thousand Oaks, CA: Sage Publications.

Educational Testing Service (2002). *ETS Standards for Quality and Fairness*. Princeton, NJ: Educational Testing Service.

Haertel, E.H., & Lorie, W.H. (2004). Validating standards-based test score interpretations, *Measurement*, 2(2), 61-103.

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

Kane, M.T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38, 319-342.

Kane, M.T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice*, 21, 31-41.

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

Messick, S. (1989). Validity. In R.L. Linn (Ed.). *Educational measurement* (3rd ed., pp. 13-103). NY: American Council on Education and Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13-23.

Mislevy, R.J. (2009). Validity from the perspective of model-based reasoning. In R.W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications*. (pp. 83-108). Charlotte, NC: Information Age Publishing, Inc.

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3-67.

NAGB (2014). *Technology and engineering literacy framework for the 2014 National Assessment of Educational Progress*. Washington, DC: National Assessment Governing Board.

Porter, A.C., & Smithson, J.L. (2001). *Defining, developing, and using curriculum indicators*. CPRE Research Report Series. Philadelphia: Consortium for Policy Research in Education.

Toulmin, S.E. (2003). *The uses of argument.* (updated ed.). Cambridge: Cambridge University Press.

Webb, N.L. (1997). Determing alignment of expectations and assessments in mathematics and science education. *National Institute for Science Education, 1*(2), 3-10.

Zeiky, M.J., Perie, M., & Livingston, S.A. (2008), *Cutscores: A Manual for Setting Standards of Performance on Educational and Occupational Tests*. Princeton, NJ: Educational Testing Service.

Zelman, M., Shmis, T., Avdeeva, S., Vasiliev, K., & Froumin, I. (2012). *International Comparison of Information Literacy in Digital Environments.* Unpublished manuscript.

Appendix A
Template for the ICTLA Validity Audit/Review

Questions related to underlying evidence:

- Is the evidence produced by each task or collection of tasks sufficient to support the claims or inferences drawn from the test scores?
- Does the evidence support the claims of the test developers?
- Is the evidence at the task- and subscore level adequate to support the test developers' claims?

Questions related to the ICT Tasks:

- Do the features of the tasks support the type of evidence intended?
- Do the tasks enable the test taker to demonstrate command of the construct-relevant knowledge, skills and abilities assessed through the ICT?
- Are there features of the tasks that are construct-irrelevant and interfere with or compete with the collection of sound evidence?

Reviewers will be asked to judge the ICT instrument and items using the following criteria:

- Are the items/ tasks technically accurate (are the items editorially and factually accurate)?
- Are the test specifications well aligned with the target construct (do the items reflect the intended test specifications at the construct level)?
- Are the tasks clear, precise and unambiguous?
- Are the tasks presented in the appropriate balance (i.e. do they adequately reflect the original test specifications)
- Is there an appropriate spread and an appropriate distribution of difficulty and complexity of the difficulty of the tasks?
- Are there sufficient items to produce a robust scale?
- Are the proposed scaling and equating processes sound?
- Will the tests support the range and depth of reports required from a robust ICT literacy assessment?
- Does the overall ICT literacy program design reflect good psychometric practice?

Auditors will also be asked to assess the adequacy of the ICT literacy design document according to the following criteria:

- Is the purpose of ICT literacy assessment clearly articulated in the ICT literacy Framework document?
- Are the scoring processes described appropriate for the stated purposes?
- Are the measurement methods described appropriate for these purposes?
- Will reliable and valid comparisons over time be facilitated?
- Are the score reports likely to be comprehensive and understood by test users?
- Will reporting be accessible to all the different stakeholders, including teachers and schools?