PEL Pensamiento Educativo. Revista de Investigación Educacional Latinoamericana
2017, 54(2), 1-24

Pontificia Universidad Católica de Chile
Facultad de Educación

Empowering End Users to Evaluate Score Reports: Current Challenges and Proposed Solution

Alentando los distintos actores a evaluar informes de resultados de pruebas estandarizadas: desafíos y propuestas

Fernanda Gándara & Francis Rick
The University of Massachusetts, Amherst, United States

Abstract

Score reporting is one of the most critical steps to the valid use of test scores, yet end users are not necessarily aware of their relevance. Importantly, end users do not always possess the knowledge or instruments to evaluate the quality of a score report, as there is no explicit guidance in the literature that empowers them to do so. Score reports are sometimes accompanied with interpretive guides, that allows stakeholders to make better sense of the data, but that do not enable end users to develop independent and critical evaluations of the reports that they receive. To that end, in this study we analyze the evaluation form provided in the Hambleton and Zenisky (HZ) model for developing score reports, to understand to what extent this form is clear, useful, and meaningful to evaluate these across different contexts. Using a small scale focus group, we were able to document the main problems with the HZ form to evaluate score reports. We were also able to identify appropriate directions and modifications to the form, to ultimately conduct subsequent studies that allow us to develop an end-user evaluation for to assess the quality of score reports.

Keywords: score reporting, testing, validity, educational measurement

Post to:
Fernanda Gándara
Email: fergandara@gmail.com

ISSN:0719-0409 DDI:203.262, Santiago, Chile
doi: 10.7764/PEL.54.2.2017.11

Resumen

El informe de resultados es uno de los pasos más críticos para el uso válido de pruebas estandarizadas, aun así, los usuarios de estas no están necesariamente conscientes de su relevancia. Es importante notar que los destinatarios no siempre poseen el conocimiento o los instrumentos para evaluar la calidad de un informe de resultados, ya que no hay una guía explícita en la literatura que los alienta a hacerlo. Los informes de resultados son algunas veces acompañados con guías de interpretación, que permite a los participantes interpretar los datos, pero que no les permite desarrollar una evaluación independiente y crítica de las evaluaciones recibidas en los informes. En este estudio analizamos la forma de evaluación proporcionada en el modelo de Hambleton y Zenisky (HZ) para el desarrollo de informes de resultados para entender hasta qué punto el formulario resulta claro, útil y valioso para evaluaciones dentro de distintos contextos. Se utilizó un grupo focal a pequeña escala donde se documentó los principales problemas con el formulario HZ para la evaluación de informes de resultados. Se pudo identificar modificaciones y directrices apropiadas para el formulario, para conducir, finalmente, estudios subsecuentes que permitan desarrollar una evaluación por parte de los usuarios para examinar la calidad de los informes de evaluación.

Palabras clave: informes de evaluación, evaluación, validación, medición educacional

Reporting assessment results is one of the most challenging aspects of test development (Zenisky & Hambleton, 2012). Score reports convey a message, typically between testing agencies and targeted groups of stakeholders. Distinct stakeholders—or relevant users of test data—hold different preferences when it comes to receiving test data (Jaeger, 2003; Zwick, Zapata-Rivera, & Hegarty, 2014), so a major challenge is to create reports that meet the requirements of different groups. Another challenge comes from the fact that the content of these reports may be of different types: the message can be summative, diagnostic, or normative in nature (Hambleton & Zenisky, 2013). For each type of reporting (summative, diagnostic, or normative), agencies have to make decisions about which type of scores to display, about which additional information to include, and about format and design issues. And these decisions should be aligned to the intended interpretations and uses of the test scores. In addition, score reports may be delivered in different mediums. Traditionally, paper score reports were the norm, testing programs are increasingly delivering their reports via the Internet (Zenisky & Hambleton, 2012). Additional challenges emerge from online score reporting. For example, online reports have to consider the extent to which the reports will be user-interactive as well as whether and how to provide interpretive materials (Zenisky & Hambleton, 2012). Because of the large number of decisions involved and the multiple possibilities to deal with them, both paper and online score reporting practices vary substantially across agencies (Goodman & Hambleton, 2004; Knupp & Ansley, 2008; Faulkner-Bond, Shin, Wang, & Zenisky, 2013).

Many testing professionals leave the task of creating score reports to the end of the test development process and do not pay enough attention to the approach that is followed (Zenisky, Hambleton, & Sireci, 2009). In turn, those who pay careful attention to the process note that there are many decisions to be made, and therefore, developing score reports may appear as an overwhelming task. Fortunately, the psychometric literature offers guidance in the form of models for developing score reports, standards developed by professional associations, and best-practices and/or practices to avoid in the development of score reports. First, the literature offers models for developing effective score reports, namely, those by Zapata-Rivera (2011) and Hambleton and Zenisky (2013). These models share some essential characteristics. Both models are research

based. Both models stress the importance of targeting specific audiences and provide similar recommendations in terms of how to address their needs. Both models are based upon the idea of a pipeline where reports are initially crafted, later prototyped internally and externally, and finally tuned so as to produce score reports that are adequate and useful (Hambleton & Zenisky, 2013; Zapata-Rivera, 2011; Zenisky & Hambleton, 2015). In relation to their differences, only Hambleton and Zenisky (2013) addresses issues of monitoring and revising/redesigning score reports. Both models are similar, but Hambleton and Zenisky's model is more comprehensive and provides more detail to guide the work of test developers.

Second, professional associations related to testing provide standards to foster the appropriate development and use of tests. The most prominent set is the *Standards for Educational and Psychological Testing* (hereafter referred to as *Standards*; American Educational Research Association [AERA], American Psychological Association [APA], & National Council of Measurement in Education [NCME], 2014). The *Standards* provide at least 13 guidelines that directly affect the process of score reporting. Broadly, these standards can be classified into three categories: (a) those that refer to the process of developing score reports, (b) those that refer to the responsibilities that testing programs have in relation to the interpretation of the reports, and (c) those that refer to the content of the score reports. More details are presented in Table 1.

Table 1
Standards Related to Score Reporting from the Standards for Educational and Psychological Testing

Topic	ID	Content
Process	6.0	To support useful interpretations of score results, assessments instruments should have established procedures for . . . reporting. Those responsible for . . . reporting . . . should have sufficient training and supports to help them follow the established procedures. Adherence to the established procedures should be monitored, and any material errors should be documented and, if possible, corrected.
	6.13	When a material error is found in test scores or other important information issued by a testing organization or other institution, this information and a corrected score report should be distributed as soon as practicable to all known recipients who might otherwise use the erroneous scores as a basis for decision making. The corrected report should be labeled as such. What was done to correct the reports should be documented. The reason for the corrected score report should be made clear to the recipients of the report.
	9.16	Unless circumstances clearly require that test results be withheld, a test user is obligated to provide a timely report of the results to the test taker and others entitled to receive this information.
	9.20	In situations where test results are shared with the public, test users should formulate and share the established policy regarding the release of the results (e.g., timeliness, amount of detail) and apply that policy consistently over time.

Responsibilities of testing programs with regards to interpretation	6.10	When test score information is released, those responsible for testing programs should provide interpretations appropriate to the audience. The interpretations should describe in simple language what the test covers, what scores represent, the precision/reliability of the scores, and how scores are intended to be used.
	12.15	Those responsible for educational testing programs should take appropriate steps to verify that the individuals who interpret the test results to make decisions within the school context are qualified to do so or are assisted by and consult with persons who are so qualified
Content	2.4	When a test score interpretation emphasizes differences between two observed scores of an individual or two averages of a group, reliability/ precision data, including standard errors, should be provided for such differences.
	3.17	When aggregate scores are publicly reported for relevant subgroups . . . , test users are responsible for providing evidence of comparability and for including cautionary statements whenever credible research on theory indicates that test scores may not have comparable meaning across these subgroups.
	6.11	When automatically generated interpretations of test response protocols or test performance are reported, the sources, rationale, and empirical basis for these interpretations should be available, and their limitations should be described.
	6.12	When group-level information is obtained by aggregating the results of partial tests taken by individuals, evidence of validity and reliability/ precision should be reported for the level of aggregation at which results are reported. Scores should not be reported for individuals without appropriate evidence to support the interpretations for intended uses.
	12.17	In educational settings, reports of group differences in test scores should be accompanied by relevant contextual information, where possible, to enable meaningful interpretation of the differences. Where appropriate contextual information is not available, users should be cautioned against misinterpretations.
	12.18	In educational settings, score reports should be accompanied by a clear presentation of information on how to interpret the scores, including the degree of measurement error associated with each score or classification level, and by supplementary information related to group summary scores. In addition, dates of test administration and relevant norming studies should be included in score reports.
	12.19	In educational setting, when score reports include recommendations for instructional intervention or are linked to recommended plans or materials for instruction, a rationale for and evidence to support these recommendations should be provided.

Another group of standards that guides the development of score reports was developed by the

International Test Commission (ITC). Between 2000 and 2012, the ITC developed and published four series of guidelines to enhance responsible testing practices. The *Guidelines on Quality Control in Scoring, Test Analysis and Reporting of Test Scores* (ITC, 2012) are meant to support the operation of large-scale assessments across the globe. These are guidelines—not standards—and should be adapted into and used as locally developed standards. This set of guidelines contains at least 15 that directly address the issue of reporting. Broadly, the guidelines may be grouped into four categories: (a) those related to the process, (b) those related to interpretation, (c) those related to content, and (d) guidelines related to the security of the score reports. More information is provided in Table 2.

Table 2
Standards Related to Score Reporting from the ITC Guidelines on Quality Control in Scoring, Test Analysis and Reporting of Test Scores

Topic	ID	Content
Process	1.2.1	Identify all the stakeholders in the testing process and agree who is responsible for decision making with respect to the different parts of the testing process.
	1.2.2	Determine and state the purpose or purposes of test use (e.g., selection, measuring achievement, research).
	1.2.14.	Determine which specific individuals, bodies or institutions should receive test results, ensuring compliance with legal constraints regarding data privacy.
	1.3.1.	Confirm that there are adequate resources (cost, time and personnel) available for efficient and appropriate scoring, test analysis and reporting of scores.
	1.4.3	Decide in advance on the process for dealing with cases where a mistake is discovered after scores have been released.
Interpretation	2.5.1.1.	Use focus groups of test takers or possibly "think-aloud procedures," "experimental studies," or even "one-on-one interviews" to gain information to assist in the development of comprehensible and instructive explanations of the score report and any interpretive guide.
	2.5.1.2.	Ensure that anyone who receives the scores has appropriate guidance in interpreting them, so here will be a proper understanding of test scores. Support this with evidence that the reports allow users to make defensible interpretations.
Content	1.2.13.	Agree upon the degree of detail with which scores should be reported to the test takers and institutions involved, and what additional information regarding score distributions and score use should be delivered.
	1.2.15.	Determine whether reports can or should provide other personal information (e.g., whether the test content was modified, how many items were completed, what accommodations for disabilities were offered).

	2.5.1.5.	Clarify to what level different scores can be relied upon (e.g., where sub-scores have too low reliability to use in making high-stakes decisions). The decision if subtest scores are reported should also be based on (a) the theory of a test, (b) the aim of testing and psychometric properties of the subtest scores.
Security	2.5.2.1.	Take steps to ensure that the individual score report cannot be forged by the test taker.
	2.5.2.2.	Avoid editing the institution report: editing may cause serious problems. If there is a need to change one or more scores, use allocated software, or create the report again.
	2.5.2.3.	Encrypt the electronic files of score reports for storage and transfer.
	2.5.2.4.	Ensure that score reports are only sent to appropriate individuals. Do not send score reports that are more inclusive than necessary. It may be easier to send the same complete report to all test users, but to maintain candidate confidentiality, only relevant results should be sent to each test user.
	2.5.2.5.	Inform institutions that only the report sent directly to the institution – and not the test taker's copy of the report (which can be faked) – is to be used for official purposes. Also recommend the institutions to do routine verifications on the institution report.

Lastly, researchers have identified a large list of best practices for individual score reports. Emblematic in this sense is the work of Goodman and Hambleton (2004), who studied score reports and interpretive guides across the US and Canada. By analyzing these documents with an unprecedented level of detail, the authors identified best practices and made general recommendations to improve design and content issues. Some recommendations include ensuring that score reports are clearly written, concise, and visually attractive; avoiding statistical jargon; and giving consideration to different users in the creation of the reports. Goodman and Hambleton also highlighted several problematic features. For instance, the authors found that many reports presented too much information without addressing key issues, such as the provision of measures of precision or definitions of key terms. Ryan (2006) echoed these results by stating that many score reports do not present contextual information regarding the description of scores and achievement levels, or do not explain students' performance with the appropriate degree of specificity. In response, researchers have looked at score reporting more closely and the literature currently provides a series of recommendations for developing general score reports (Goodman & Hambleton, 2004; National Education Goals Panel, 1998), reports to be disseminated online (Zenisky & Hambleton, 2012), group-level score reports (Zenisky et al., 2009), and score reports for English Learners' parents (Zapata-River et al., 2014; Faulkner-Bond, Shin, Wang, & Zenisky, 2013). It is important to stress that many of the recommendations are consistent with the *Standards* (AERA et al., 2014), and are addressed in the models by Zapata-Rivera (2011) and Hambleton and Zenisky (2013).

Despite the apparently wide consensus among researchers about what works and does not when it comes to reporting, there are many score reports that still present problems (Zapata-Rivera, 2011). One of the most critical issues around score reporting is that of interpretability. Research shows that many score reports are misinterpreted or are difficult to interpret across several audiences (Ward, Hattie, & Brown 2003; Goodman & Hambleton, 2004; Ryan, 2006; Zenisky et al., 2009; Whittaker, Williams, & Wood, 2011; van der Kleij & Eggen, 2013). Another common problem with score reports is that they do not include information regarding the precision of scores or even the purpose of the assessment (Goodman, & Hambleton, 2004). This finding is rather surprising considering the large amount of effort that professional asssociations and researchers had made to highlight the importance of being explicit about these. Score reports also tend to lack information regarding key terms, despite using abundant statistical jargon (Goodman & Hambleton, 2004). Other problems presented in literature include scarce contextual information related to descriptions of scores and achievement levels (Ryan, 2006; Zenisky et al., 2009), the lack of diagnostic information to connect assessment results with instruction (NEGP, 1998), and the lack of subgroup reporting (Zenisky et al., 2009), despite the salient important of certain group comparisons. These problems reduce the usefulness of a score report and attention needs to be paid to them, or all the resources invested into test development may be wasted.

As outlined, there are plentiful reasons why score reports remain problematic. One explanation is that test agencies do not allocate sufficient resources to reporting efforts, or do not take reporting considerations into account from the beginning of the testing cycle (Zenisky & Hambleton, 2015). Therefore, the problem could be process-related and test developers should make larger efforts to improve reporting practices. A second possibility is that developers are not aware of the extensive amount of resources available to support their reporting endeavors. The problem could be thus, knowledge-related and could be solved if research was more effectively disseminated to test developers in charge of reporting processes. A third possibility is that efforts put into score reporting are not as effective as test developers intend. For example, developers may have not figured out what truly works for the particular audiences and context of their testing program. It may be that research on the core audiences has not been appropriately conducted. Or it could be that the audiences have changed over time, so the new audiences' information needs have to be updated. In any case, a gap between the amount of information and support offered in the literature and the outcomes of current score reporting practices among developers prevails.

Evaluating the quality of score reports

A different explanation of the gap between research and practice when it comes to score reporting is rooted in the lack of external pressure and accountability on reporting decisions. Stakeholders—such as parents, teachers, school communities, or policy makers—are concerned about the quality of a test, but it is rare to find public discussion around the reports and how these affect the validity of the use of scores. Even technical documentation seems to undermine the issue of reporting compared to other aspects of testing. Technical manuals are very descriptive around issues of administration or around the properties of constructs and scores, but do not equally emphasize reporting: why does the report portray a particular set of information? Why are the format decisions of a given type? Reporting decisions are critical yet this message has not been necessarily adopted by the public.

The psychometric literature does not offer concrete support to end users to evaluate the quality of a score report, and to consequently demand for better results when it comes to reporting. Few researchers have developed report evaluation tools. In their seven-step model, Hambleton and Zenisky (2013) introduced an evaluation form to be used by developers in their attempts to finalize score reports. The HZ form is a set of questions aimed at reflecting upon the relevant aspects of the score reports. The HZ form assesses the quality of the score reports using 36 questions split across eight dimensions: (a) Overall – questions that reflect holistically on the report; (b) Content - Report Introduction and Description – questions concerning the relevant information about the assessment and/or program that should be provided at the beginning of the report (e.g. does the report explain the purpose of the assessment?); (c) Content - Scores and Performance Levels – questions pertaining to the clarity of the score scales included in the report; (d) Content - Other Performance Indicators- questions concerning the clarity of subscales or item analyses, as well as the appropriate uses that should be made out of this information; (e) Content – Other – questions concerning the support that agencies provide to end users in order to help them interpret and use scores; (f) Language – questions concerning terminology and tone used in the report; (g) Design – questions about the logic under which topics/questions are organized, as well as questions on the format decisions used in the report; and (h) Interpretive Guides and Ancillary Materials – questions pertaining the additional material used to support the effectiveness of the score reports.The development of the HZ form followed research findings and best practices in reporting: the questions and sections included in the form reflect the most relevant aspects suggested by the literature around score reports. See Appendix A for more details on the form.

Additional efforts to create instruments to evaluate score reports were put forth by Gotch and Roberts (2014). Under the conviction that systematizing the evaluation of score reports could benefit the psychometric practice (Roberts & Gotch, 2016), these authors developed and tested their own instrument, largely on the HZ form and framework. The purpose of their work was to develop a tool for researchers, to ultimately gauge the impact of the quality of score reports on the validity of test score uses (Gotch & Roberts, 2014; Roberts & Gotch, 2016). To that end, they took the HZ form and eliminated questions that felt were not relevant for researchers. They also transformed the different questions into statements, mainly to enable the use of a rating scale. In total, the authors kept 31 criteria across 6 domains.

Gotch and Roberts (2014; Roberts & Gotch, 2016) conducted two different studies to develop, evaluate, and refine their instrument. First, the authors used their form to evaluate 41 score reports used by different U.S. state testing programs. The objective was to test and finalize the form. Each of the 31 statements was accompanied by a 3-point rating scale corresponding to each the following criteria: not met (0), partially met (1), or fully met (2). The authors evaluated the 41 reports themselves and found that the GR form yielded promising results. On one hand, the exact matches produced by the form reached 67% across all the reports, and exact adjacent ratings were provided for 94% of the criteria. On the other hand, the rating scale allowed the authors to compute overall scores for each of the reports. The authors examined the reports that obtained the highest and lowest ratings, and found that those with high scores were actually of better quality than those with low scores. The authors also noted that the form was able to inform researchers about the specific variation of score reports across different areas of interest (Gotch & Roberts, 2014).

Since the purpose of the GR form was to inform research, one of its most important considerations pertained to reliability. If the reliability of using the form is low, then it cannot properly support research around the impact of score reports on the valid use of test scores. Therefore, in their second study, the authors focused on investigating the reliability properties of the GR form. Based on their previous experience, the authors made minor modifications: they used a 4-point rating scale instead of the 3-point scale, and provided specific descriptors of each score point. The authors asked 4 graduate students to evaluate 3 score reports using the modified form. Next, the authors conducted a two-facet generalizability study, where score reports were fully crossed with domains and raters. The two most important findings were that the form yielded highly reliable data (G=0.78) but that there were other unidentified systematic sources of variance influencing the assessment of the score reports (Roberts & Gotch, 2016). These results are promising, and the GR form possesses the potential to systematize the evaluation of the quality of score reports among researchers.

Purpose

The psychometric literature provides only two instruments to evaluate the quality of score reports: the HZ form, which is meant to assist test developers, and the GR form, which is meant to assist researchers. The HZ form was developed to help systematize and improve the process by which score reports are created (Hambleton & Zenisky, 2013; Zenisky & Hambleton, 2015). In turn, the GR form was developed to enable a systematic analysis of score reports and their impact on validity considerations (Gotch & Roberts, 2014; Roberts & Gotch, 2016). One of the indirect consequences of the GR form is to empower the research community, as researchers can now use the form to encourage and demand for better score reports. We believe this is a valuable contribution to the field, but that this contribution could be better capitalized if we could empower end users—e.g. parents, teachers, and school communities—to evaluate the quality of reports and consequently, to present similar requests to test developers. We entirely agree with the sentiment that score reports are critical to the valid use and interpretation of test scores (Hattie, 2009; Roberts &Gotch, 2016), but we perceive this idea as being largely absent from the general public' discussion when it comes to evaluating tests and their impact. While there is abundant literature around the topic of score reports, research has not produced tangible tools that educate stakeholders and empower them to demand better quality reports. The purpose of this study is to take a first step in that direction.

The Hambleton and Zenisky (2013) evaluation form is a good starting point to develop such tool. The HZ form is research-based, comprehensive, and focuses on aspects that are certainly relevant to many stakeholders. As such, we agree with Gotch and Roberts (2014) that the HZ form serves as a basis to develop other evaluation forms, including a form to educate and empower end users. The specific purpose of this paper is evaluate the appropriateness and usefulness of the HZ form—as originally conceived—to evaluate finalized score reports from the perspective of end users. To that end, we examined the clarity, usability, and meaningfulness of the HZ form to evaluate a typical individual-level score report. Our findings, together with the findings from similar studies, will inform two consecutive studies meant to develop a new evaluation form so that end users can evaluate score reports independently and in a wide variety of contexts.

Method

To evaluate the clarity, usability, and meaningfulness of the HZ form as a score report evaluation tool, we conducted a small-scale focus group study using a purposeful sample.

Participants

Participants were six graduate students (three men, three women) who volunteered to take part in a focus groups. All participants attended a School of Education in an American university and were, therefore, relatively familiar with the essential concepts around educational assessments. Furthermore, based on answers to a brief demographic survey, all participants remembered receiving at least one score report describing their own performance on a standardized assessment, and had seen an average of 3.5 unique score reports in the last five years.

Materials

Participants evaluated a sample score report using a paper version of the HZ form (see Appendix A). The form was formatted as a two-column table, where questions were presented in the first column and blank spaces for answers were presented in the second column.

The score report (see Appendix C), which displayed a fictitious student's performance on a high school-level English assessment, was selected from a collection of score report samples obtained from department of education websites for over 25 states. A meticulous selection process was followed to ensure that idiosyncrasies of whichever report was used would not play an undue role in the evaluation of the HZ form. The report was ultimately selected because it represents an "average quality" report based on the authors' holistic judgment and includes key content and design elements that are increasingly common in modern score reports, such as sub-area performance indicators, a graphical display of overall performance, and interpretive text for the overall score and subarea scores. (Please note that identifying information has been blurred or omitted in the figure presented in Appendix C, but the version of the score report used in the focus groups was unaltered).

Procedure

Data collection. We assigned the participants to two groups of three, and asked each group to attend a one-hour long focus group. At each meeting, the participants engaged in a 20-minute exercise, where they had to use the HZ form to evaluate a sample score report. After using the form, the participants engaged in a 30-minute discussion around their experience using it. The discussion was semi-structured and meant to gather feedback on issues that are central to this study. The focus groups' discussions were recorded (with participants' consent) and transcribed. More details about the focus group format and questions are provided in Appendix B.

Data analysis. First, we conducted a qualitative analysis of the focus group transcripts. The purpose was to identify those themes that participants considered most relevant, and to summarize their perspective on each. This procedure did not involve pre-defined themes. Next, we analyzed

the responses that the participants gave to the questions in the HZ form. We analyzed this data to understand the extent to which the HZ form was clear, usable, and meaningful to participants in their attempt to evaluate the score report provided. In particular, based on their responses, we computed several frequencies on the following: (a) questions that were confusing to the participants, (b) questions that elicited consistent responses among the participants, (c) questions that could have been worded as rating scale instead of open ended questions, (d) questions that the participants could not respond because they needed follow-up activities, and (d) questions that were redundant to participants. For instance, to inform our evaluation of the clarity of the HZ form, one of the frequencies we computed indicated how many questions were perceived as confusing based on responses that included terms like "I am not sure" and "I do not understand."

Results

Qualitative analysis of the transcripts

The first piece of the analyses consisted on inspecting the focus groups transcripts. Several topics emerged, as we explain below.

A first issue that emerged in the focus group discussion was related to the uncertainty that participants had regarding the stakeholders (1): who are we evaluating this for? And who is going to use this form? Without this information, it was impossible to answer some of the questions from the HZ form, such as question I.B, which reads *"Does the score report reflect the reporting interests and informational needs of key stakeholders?"* Without knowing who the stakeholders were, these questions were ambiguous and more generally, participants felt that they could not put their answers into context.

> "When I saw the question about stakeholders, I thought that "this is a family score report", so I thought about the stakeholder as just a parent or a test takes, but usually stakeholders also include policymakers or teachers, so I was a little bit confused with that question."

The HZ form was meant to inform test developers who would know who the stakeholders were. At first glance, questions referring to stakeholders seem redundant or inappropriate if the purpose of the form we want to develop is to inform stakeholders themselves. Yet at this stage we wanted to avoid making any adjustments based on our own judgment and get a sense of how the participants reacted to each question.

A second issue that emerged is related to the image that participants created of the ideal score report (2) as suggested by the HZ form. Based on the questions, the participants imagined a cluttered and information-dense report as an ideal. Interestingly, this is exactly what Goodman and Hambleton (2004) warned against; a well-accepted recommendation is that score reports should be concise and uncluttered. Therefore, parts of the form led to the wrong impression of what good score reports look like, although most suggestions are in line with what researchers suggest constitutes a good score report.

A third issue that emerged from the focus group discussion was that participants thought that the content of the questions was inappropriate (3). First, participants mentioned that many questions were redundant and that they felt as if they were providing the same answer multiple times. Another set of comments had to do with the importance of the questions, as participants felt that some were not very important (although, according to the participants, fully determining that required acknowledging the purpose of the evaluation of the score report). Also, participants felt that some important questions were left out of the HZ form, such as questions about graphical elements of reports. The report evaluated by the participants (see Appendix D) had one graphic and the participants had numerous comments about it. In particular, they felt that the graphic was confusing (e.g. the scale was not clear) and that they could not express this using the HZ form.

Related to this issue, participants also mentioned that the language of the questions seemed to be old-fashioned. In addition, they noted that some questions were very easy while others required expertise of some sort. In other words, they felt that questions varied in terms of difficulty. Also, some questions could not be answered without additional information (e.g. *"Is there an interpretive guide prepared, and if so, is it informative and clearly written?"*). Participants were a little bit confused about whether they had to "figure out" some of this missing information from the report or not.

> "Apart from the question about the stakeholders, I think there is also a question about 'reports or materials are available in different mediums, does it align to related materials published' those questions, I don't know how you can answer just based on this, so there is need for more information. And then, by putting the school average, the district average, and the state average, I don't know if they were trying to lead us to say which are the stakeholders."

A fourth issue that emerged in the conversation was that certain questions catalyzed meaningful reflections about the score report (4), something that was perceived as positive. This is an interesting finding, as we want to develop a form that invites stakeholders to think critically about reports.

> "I do think that there were a couple of prompting questions that were useful in some ways. I hadn't really processed the scale of the marker, so when I was asked about that, I was like 'oh yeah—that doesn't make sense!"

The fifth issue that emerged from the discussion was related to the format of the questions in the HZ form: participants thought that the format of the answer options was not always appropriate (5). The original HZ form is a set of questions without any scale attached to them. For the purpose of this study, we added a blank space next to each question so that participants could provide their responses in whatever format they wanted. Some participants complained about the size of the answer box not being very big, which is not a problem of the HZ form but of our print since the size of the answer box was a result of our printing decisions. Yet, regardless of the concrete size of the answer boxes, some participants felt that the response expectations were not clear, that some questions could be answered with a single word (e.g. "yes" or "no") and that others required further explanation, despite the fact that all questions were formatted equally.

> "I felt like I didn't really know how much they wanted to know. So the first question, I couldn't finish because the box wasn't big enough, but other questions made me wonder if I just had to write yes or no. Do they want a super long answer or not?"

Participants suggested using a Likert scale for some of the questions, which would make the experience of using the form easier, and potentially more useful. In addition, participants mentioned that the subjective nature of the questions suggested answer options of the type "I feel like yes" instead of more objective answer options such as "yes/no".

A sixth issue that emerged was that there were problems with the structure of the form (6). Some suggestions emerged in this regard. A first suggestion made by the participants was to change the order of the questions, and begin with the simple questions (e.g. those around design issues) and to finish with the overall/most comprehensive/complex ones. Leaving the summative questions for the end, after having reflected upon all the specific issues of the form, made more sense. A second suggestion was to get rid of the headers in between the different sets of questions, as they took space and were not attended to. A third suggestion was to include a table of contents at the beginning of the form, so as to give participants a sense of the topics and questions that are included. This way, users can plan their responses accordingly.

> "…I was thinking that I wish he had had a table of contents—like "these are the questions we'll go through", because then instead of writing the same thing so many times, I would have saved some of my responses for the most relevant [section]."

A seventh theme that emerged had to do with the clarity of the questions, as participants stated that some questions were confusing (7). According to them, the confusion emerged from the word choice: participants suggested that using simpler words would solve the problem.

The last issue that emerged from the focus groups was that the HZ form, in its current version, is not entirely useful to provide an overall rating of a score report (8). On one hand, without rating scale answer options, it is impossible to provide a summative score from which to make a final judgment. Even with rating scale answer options, it is difficult to combine the questions from different sections. On the other hand, using this form requires some type of specialization (e.g. psychometrics' knowledge): it is unlikely that one person alone could use the form to judge the quality of a score report. Participants did not minimize how relevant this form could be for evaluation purposes, but felt that it was built with a particular report in mind, which that was not necessarily assisting end users to evaluate reports.

> "To me it felt like the checklist was created with a specific score report in mind, and I bet that if we were looking at that score report, some of these questions would make a lot more sense."

Analysis of participants' responses

In addition to analyzing the focus group transcripts, we analyzed the responses given by the participants to the HZ form, with an emphasis on the clarity, usefulness, and meaningfulness of the form.

On the clarity of the HZ form. ***To what extent were the questions and sections in the HZ form clear to users?*** To respond to this, we looked at the number of questions that respondents

found confusing, as per two criteria: the questions that had answers of the type "I do not understand the question", and questions that had answers that indicated that the participants were. indeed, confused. We added these two up to compute the frequency of answers that were confusing to participants, per question. We defined three levels of confusion: not confusing – questions that did not confuse participants; little confusion – questions where we found 1-3 confusing answers (out of 6); and confusing – questions were confusing answers were 4 or more. Overall, questions presented no confusion (78%) or little confusion (19%), as seen in Figure 1.

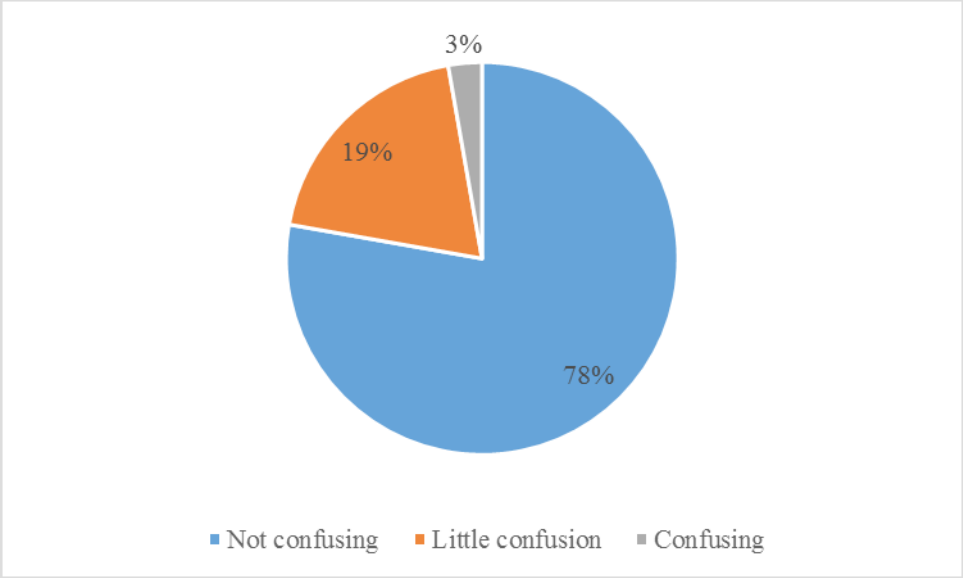

Figure 1. Questions that were confusing to participants

To what extent did the HZ form elicit consistent answers? To respond to this, we examined the degree to which each question elicited similar conclusions from participants, where answers such as "yes", "I think so", and "mostly" were treated as representing the same conclusion. The same rule was applied to expressions signaling negative conclusions (e.g., "no", "I don't think so", "not at all"). The sets of responses for each question were rated as consistent if all participants had the same conclusion; partially consistent if at least four out of six participants had the same conclusion; and inconsistent if fewer than four participants had the same conclusion. As shown in Figure 2, 6 questions were rated as eliciting consistent conclusions, 16 as partially consistent, and 14 as inconsistent.

It is worth noting that the consistency of responses is influenced both by characteristics of the questions and characteristics of the participants. A different group of evaluators might provide a much more—or less—consistent set of responses. Furthermore, while consistency might be practical when making a summative assertion about score report quality, a certain level of inconsistency means that the HZ form succeeds at prompting different perspectives. Inspection of responses rated as partially consistent or inconsistent revealed that the following question types were more prone to inconsistent responses: questions that are very open-ended ("What are your *overall impressions* of the report?"); require knowing the opinion of other individuals ("Does the score report reflect the reporting interests of *key stakeholders*?"); ask about multiple aspects within the same question (Does

the report provide t*elephone numbers, website addresses, or mailing addresses* to which questions can be directed?); or are perceived as confusing ("Is there information describing the *unit of analysis* being reported?").

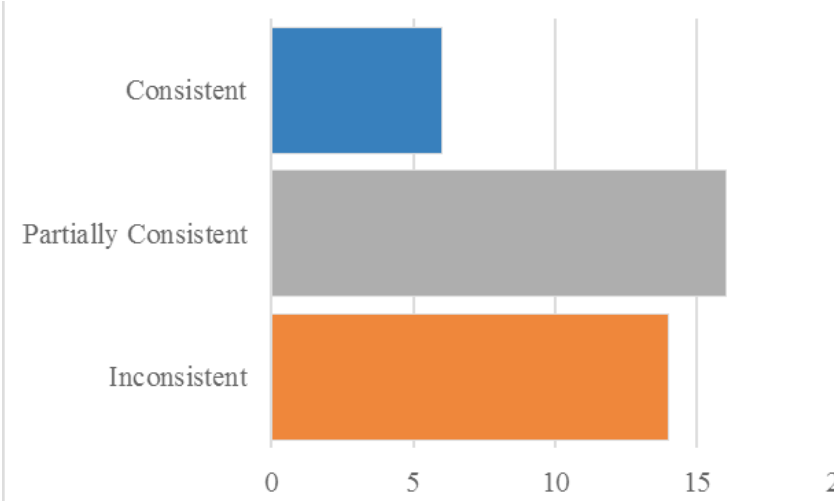

Figure 2. Consistency ratings for 36 questions

On the usability of the HZ form. ***To what extent is the format of the questions in the HZ form appropriate?*** One of the issues that was raised in the focus group discussion was the appropriateness of the format of the questions. This issue was also raised by Gotch and Roberts (2014), who ended up using a fixed scale to accompany the statements in their form. To understand whether this issue pertained to few, most, or all of the questions, we created frequencies of those answers that could be treated as yes/no answers (see Figure 3) or as rating scale answers, more generally (see Figure 4). As observed, few questions (24%) were answered as yes/no questions, but most of the questions (58%) were answered as with a rating scale, suggesting that most questions could benefit by using some fixed scale.

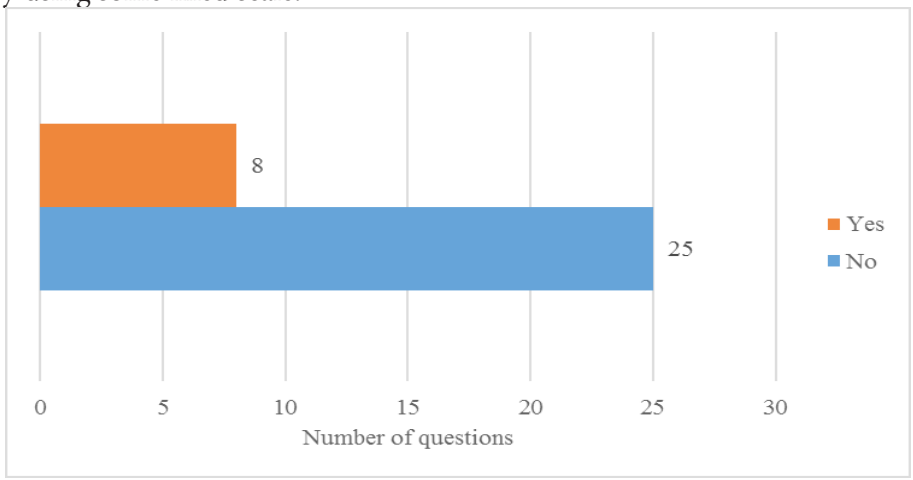

Figure 3. Questions answered as yes/no

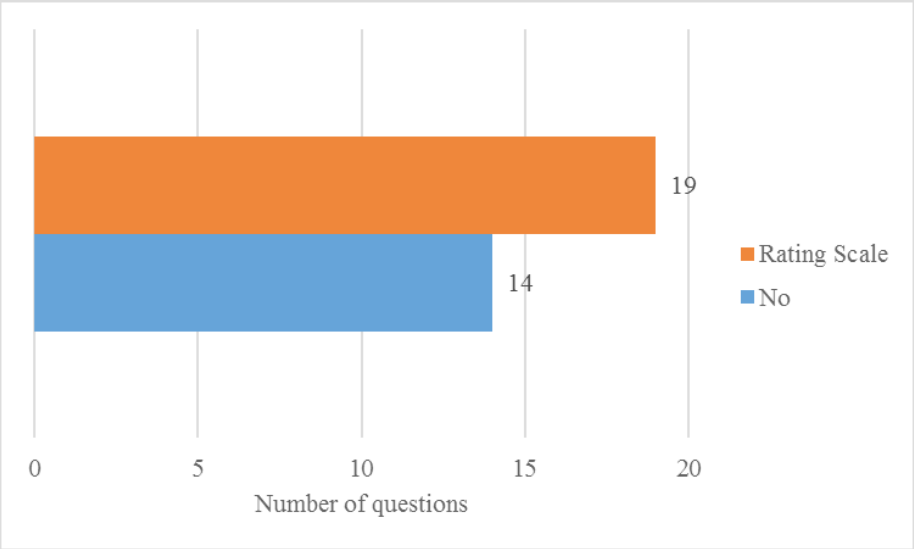

Figure 4. Questions answered as rating scale answers

To what extent did the questions in the HZ form require statistical or design expertise on the side of evaluators? The participants had the option to omit the responses to some of the questions if they felt that they required expertise of some sort (e.g. statistical knowledge or design knowledge). As such, we were able to create frequencies of questions that required technical expertise. As shown in Figure 5, the participants felt that only 14% of the questions required some expertise.

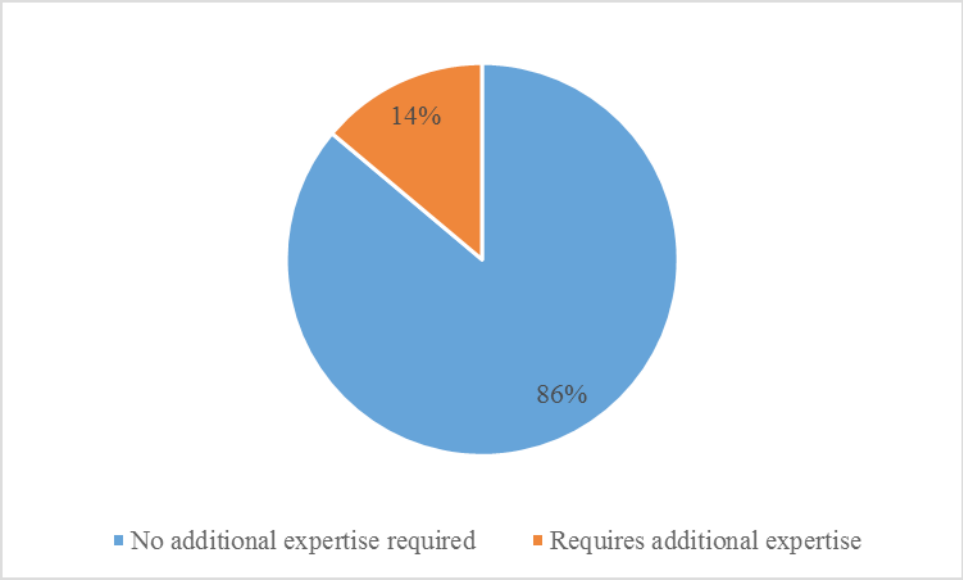

Figure 5. Level of expertise required by the questions

On the meaningfulness of the HZ form. ***To what extent did the questions in the HZ form require follow-up activities?*** The participants also had the option to omit the responses to some of the questions if they felt that providing a proper answer required some type of follow-up inquiry or activity. A simple frequency analysis of their answers indicates that 64% of the questions did not

require any type of follow-up activity, but that 25% of them did require some level of follow-up activity (between 1 and 3 participants thought this was the case), and that 11% of the questions required a high level of follow-up activity (between 4 and 6 participants thought this was the case).

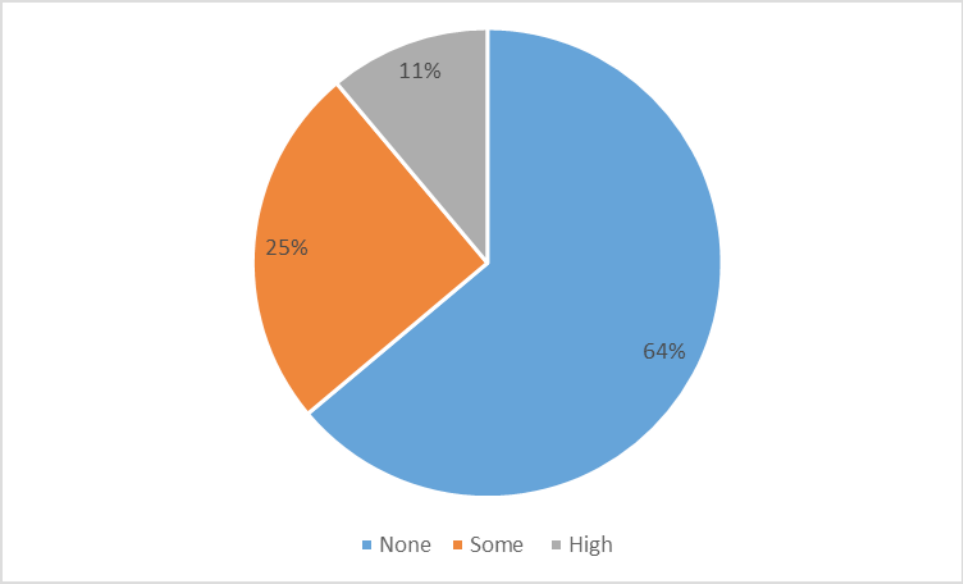

Figure 6. Level of follow-up activities

To what extent were the questions in the HZ form redundant? While participants mentioned that the form had redundant questions, based on their answers, only two questions (6%) could be classified as redundant. For these two questions, some participants used answers that began with "see above" or "again…", yet such answers were mostly absent.

Discussion

Score reports are critical to the valid use of test scores, as they are the only medium that connects test developers with stakeholders. If stakeholders are not able to understand or use the information in a score report, all the efforts and resources put into test development and data collection are simply wasted. Thus, during the past 10 years, researchers have importantly increased their focus on reporting, and have put forth multiple resources to improve reporting processes. Consequently, and since then, reporting practices have advanced, especially when it comes to the process of developing reports. Research has positively and effectively impacted psychometric practice.

Despite all the progress made, end users seem to not be aware of how critical score reports are and to what extent they should demand more clear, useful, and meaningful reports. Public discussion around testing rarely centers around issues of reporting, because knowledge of psychometric and performance data is wrongly taken as a given. Yet interpreting psychometric data is a complex process that should not be undermined. While we acknowledge that test developers care about stakeholders and implement several rounds of feedback to improve the final reports, we believe that this is not enough. There are many contexts where reports are still left to the end of test development, with little care in their production. Additionally, reports may still be unclear or of little use to many

stakeholders, who vary substantially in their assessment literacy. Stakeholders do deserve good quality score reports and need to be aware of their importance as well as of the qualities that characterize good reports. In that regard, research has fallen behind.

An accessible tool for end users to judge the quality of score reports is much needed to allow end users (e.g., parents, teachers, or policy analysts) get a sense of how good or bad score reports are. In this study, we took a first step in that direction by evaluating the clarity, usefulness, and meaningfulness of the Hambleton and Zenisky (2013) form. The HZ form was not envisioned for end users, but it constitutes a strong base from where to start this work. In line, we collected feedback from six graduate students to understand to what extent the HZ form is useful to evaluate a typical score report and what modifications should follow. Based on their responses and discussion, we can define a series of next steps to produce a form that would allow end users to gauge the quality of score reports.

Our first conclusion is that the content of the HZ form is not appropriate for end users. Based on the focus groups' transcripts, many questions felt redundant to participants, similar to what Gotch and Roberts (2014) noted in their study. The content of some questions was also inappropriate because of language issues (i.e., old-fashioned or overly complicated terms). Participants also mentioned that the questions failed to acknowledge some relevant issues about the report, and some were seen as irrelevant. These findings are not necessarily visible from participants' written responses, as only 6% of the questions were classified as irrelevant and 28% as somewhat confusing. However, their responses indicate that only 17% of the questions elicited consistent responses, meaning that participants tended to interpret questions quite differently. All of this information indicates that questions need to be improved in terms of language, redundancy, and clarity, but that end users may also require more guidance.

The issue of guidance relates to another complaint made by the participants—that the questions should have scales (or that they should not be all open ended). Overall, 58% of the questions could have been easily worded as rating scale questions, which would have made the evaluation exercise more practical. Participants felt that having ratings would have allowed them to make better conclusions about the quality of the report, which Gotch and Roberts (2014) also noted. However, this is not a recommendation that can be directly implemented in an instrument for end users: end users do not want to compare score reports across programs or make consistent ratings; rather, end users want to understand which aspects of the reports they need to pay attention to and, from them, understand what constitutes a good report in a particular context. We believe that using rating scales could be beneficial for questions that demand declarative types of knowledge, but that the more holistic questions should be left open-ended. The purpose of an end user evaluation tool would be to educate stakeholders by including a set of questions that point to aspects that are relevant in any score report, but also to empower them by allowing stakeholders to develop their own and independent judgment about score reports. We believe that a mix of close- and open-ended questions would effectively meet both purposes.

However, the ordering of the questions is relevant. One of the most interesting suggestions that came from our focus groups was that there should be a progression in terms of difficulty of

the questions: they should be ordered from easy to hard. This makes sense as some questions in the HZ form stimulate reflections about the reports that can be better capitalized in a final judgment provided at the very end. However, some participants mentioned that the image of the ideal score report that this form suggested was ambiguous and, at times, inappropriate. Whatever questions or statements are left in a final tool need to clearly point out the relevant issues in a score report. In addition, to enhance the meaningfulness of this form, it is critical to simplify questions that currently seem to require expertise of some sort (14%) and eliminate questions that require follow-up activities (36%), except when these activities speak directly to end users.

The step that immediately follows this study is to prototype an adapted instrument. Initially, we propose to modify the HZ form to include a table of contents; modify the current section headers; change the order of questions so that the most holistic ones come at the end; eliminate questions that are redundant or irrelevant to end users; reword the questions as clear statements; use rating scales when appropriate; provide more guidance for each question (include examples or use clearly defined rating scales); and make sure that the questions point towards the correct image of what good score reports look like. However, we envision several possible formats for this new tool and extensive additional feedback would be necessary. Indeed, we need to collect feedback from a larger and more varied pool of participants, who would evaluate a larger set of reports. We acknowledge that the biggest limitations of the current study are related to its breadth: we used a small pool of participants who may not be representative of our idea of end user, and who worked with a single score report. However, this limitation is not critical at this stage, since this is the first step of a rather large research agenda.

The original article was received on December 27th, 2016
The article was accepted on October 30th, 2017

References

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

Faulkner-Bond, M., Shin, M., Wang, X., Zenisky, A., & Moyer, E. (2013, April). Score reports for English proficiency assessments: Current practices and future directions. Paper presented at the annual conference of the National Council on Measurement in Education, San Francisco, CA.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education*, 17(2), 145-220. doi: 10.1207/s15324818ame1702_3

Hambleton, R. K., & Zenisky, A. L. (2013). Reporting test scores in more meaningful ways: Some new findings, research methods, and guidelines for score report design. In K. F. Geisinger (Ed.), *American Psychological Association Handbook of Testing and Assessment in Psychology* (pp. 479-494). doi: 10.1037/14049-023

International Test Commission (ITC). (2000). *International guidelines for test use.* Retrieved from http://www.intestcom.org/guidelines/index.php: International Test Commission (ITC).

International Test Commission (ITC). (2012). *ITC Guidelines for quality control in scoring, test analysis, and reporting of test scores.* Retrieved from http://www.intestcom.org/guidelines/index.php

Jaeger, R. M. (2003). NAEP validity studies: Reporting the results of the National Assessment of Educational Progress (Working Paper 2003-11). Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Knupp, T., & Ansley, T. (2008, March). *Online, state-specific assessment score reports and interpretive guides*. Paper presented at the meeting of the National Council on Measurement in Education, New York, NY.

National Education Goals Panel (NEGP). (1998). Talking about tests: An idea book for state leaders. Washington, DC: U.S. Government Printing Office.

Ryan, J. M. (2006). Practices, issues, and trends in student test score reporting. In S. W. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 677-710). Mahwah, NJ: Lawrence Erlbaum Associates.

van der Kleij, F. M., & Eggen, T. J. H. M. (2013). Interpretation of the score reports from the Computer Program LOVS by teachers, internal support teachers and principals. *Studies in Educational Evaluation, 39*, 144-152. doi: 10.1016/j.stueduc.2013.04.002.

Ward, L., Hattie, J. A. C., & Brown, G.T. (2003). The evaluation of asTTle in schools: The power of professional development. AsTTle technical report 35, University of Auckland/New Zealand Ministry of Education.

Whittaker, T. A., Williams, N. J., & Wood, B. D. (2011). Do examinees understand score reports for alternate methods of scoring computer based tests? *Educational Assessment, 16*(2), 69-89.

Zapata-Rivera, D. (2011). Designing and evaluating score reports for particular audiences. In D. Zapata-Rivera & R. Zwick (Eds.), *Test Score Reporting: Perspectives From the ETS Score Reporting Conference*. (ETS Research Report No. RR-11-45). Princeton, NJ.

Zapata-Rivera, J. D., & Katz, I. R. (2014). Keeping your audience in mind: Applying audience analysis to the design of interactive score reports. *Assessment In Education: Principles, Policy & Practice, 21*(4), 442-463.

Zenisky, A. L., & Hambleton, R. K. (2012). Developing test score reports that work: The process and best practices for effective communication. *Educational Measurement: Issues and Practice, 31*(2), 21-26.

Zenisky, A. L., & Hambleton, R. K. (2015). Good practices in score reporting. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 585-602). New York, NY: Routledge.

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2009). Getting the message out: An evaluation of NAEP score reporting practices with implications for disseminating test results. *Applied Measurement in Education, 22*(4), 359–375. doi:10.1080/08957340903221667

Zwick, R., Zapata-Rivera, D., & Hegarty, M. (2014). Comparing graphical and verbal representations of neasurement error in test score reports. *Educational Assessment, 19*(2), 116-138. doi: 10.1080/10627197.2014.903653

Appendix

A. Hambleton and Zenisky Evaluation Form (as in Hambleton & Zenisky, 2013)

Report Element		Score Report Review Questions
I. Overall	A.	What are the overall impressions of the report?
	B.	Does the score report reflect the reporting interests and informational needs of key stakeholders?
II. Content - Report Introduction and Description	A.	Does the report have a title clearly identifying what it is?
	B.	Are details provided about the content of the test(s) being reported?
	C.	Is there information describing the unit of analysis being reported?
	D.	Are the purpose(s) of the test described?
	E.	If present, does the introductory statement from the sponsoring agency (e.g., governor, commission, president, psychologist, etc.) set a positive tone for the report?
III. Content - Scores and Performance Levels	A.	Is the range of the score scale communicated?
	B.	Are the performance categories or psychological states being used (e.g., failing, basic, proficient, advanced, passing) described sufficiently for the intended audience?
	C.	Is information provided for how all of the numerical scores and classifications should be used and should not be used?
	D.	Are concrete examples provided for the use of the test score information?
	E.	Is the topic of score imprecision handled for each score that is reported? Descriptions, graphics, or numbers are all possibilities.
	F.	Have "probabilities" or "conditional probabilities" been avoided? If they are used, is the explanation clear?
	G.	Have footnotes been avoided, but if they are used, are they clearly written for the reader?
	H.	Is there sufficient information for the reader, without being overwhelming?
	I. A.	Is there any linking of test results to possible follow-up activities? For example, with educational tests, are the results linked to possible instructional follow-up?
	B.	If present, are relevant reference group comparisons reported with information on appropriate interpretations?
IV. Content - Other Performance Indicators	C.	If present, are results of performance on individual test questions reported with a key for understanding the item attributes and the performance codes?
	D.	If subscale reporting is included, are users informed about the level of score imprecision? If norms are provided, is the reference group described in sufficient detail? Are the meanings of T scores, z scores, normalized z scores, stanines, stens, percentiles, grade equivalent scores, etc. made clear?
	E.	If present, are reports of scores from other recent and relevant tests (NRTs, etc.) explained?

V. Content – Other	A.	Does the report provide telephone numbers, website addresses, or mailing addresses to which questions can be directed?
	B.	Does the report provide links to additional resources about the test, testing program, and/or understanding examinee performance?
VI. Language	A.	Is the report free of statistical and other technical jargon and symbols that may be confusing to users?
	B.	Is the text clearly written for users?
	C.	Is the report (or ancillary materials) translated/adapted into other languages? If so was the translation/adaptation carried out by more than a single person, and was an effort made to validate the translated/adapted version?
VII. Design	A.	Is the report clearly and logically divided into distinct sections to facilitate readability?
	B.	Is a highlight or summary section included to communicate the key score information?
	C.	Is the font size in the different sections suitable for the intended audience?
	D.	Are the graphics (if any) presented clearly to the intended audience?
	E.	Is there a mix of text, tables, and graphics to support and facilitate understanding of the report data and information?
	F.	Does the report look friendly and attractive to users?
	G.	Does the report have a modern "feel" to it with effective use of color and density (a good ratio between content and white space)?
	H.	Is the report free of irrelevant material and/or material that may not be necessary to address the purposes of the report?
	I.	Is the "flow" for reading the report clear to the intended audience starting with where reading should or might best begin?
	J.	Does the report align in layout and design to related materials published by the testing program?
VIII. Interpretive Guides and Ancillary Materials	A.	Is there an interpretive guide prepared, and if so, is it informative and clearly written? Has it been field-tested? Are multiple language versions available to meet the needs of intended readers?
	B.	If there is an interpretive guide, is there an explanation of both acceptable and unacceptable interpretations of the test results?

B. Focus Group Protocol

Focus Group Protocol

1. (2 minutes) Briefly describe the purpose of the project (improve an existing checklist by making it more accessible, etc.; emphasize that criticism is welcome).
2. (2 minutes) Explain the tasks participants will be performing (first use the checklist individually to evaluate a score report, then have a group discussion about their experience using the checklist).
3. (5 minutes) Go over the instructions on the first page of the worksheet together. Check if there are any questions.
4. (1 minute) Explain how the recording from the focus group will be used, and obtain permission to record the discussion.
5. (Approximately 20 minutes) Individually going over the checklist.
6. (Approximately 25 minutes) Group discussion.

Questions for the Group Discussion

Usefulness

1. Is this checklist a useful tool to evaluate the quality of a score report? Why? Why not?

Structure

2. What did you think about the sequence of items?
    - Did the order make sense?
    - Were there any items that seemed to come too soon, or too late?
3. What rating scale do you think could make sense for these questions?
4. What might be some ways to summarize how good a score report is based on all the questions? In other words, after evaluating a score report using this checklist, how could we give someone an "overall result"?
5. What did you think about the way the items were grouped (e.g., "overall", "language")?
    - Could you recommend other ways to group the items that might also be helpful, or that might be more helpful?

Content

6. Are there any questions that have confusing words, or that are overall confusing?
    - Which ones?
    - How would you change those questions to make them better?
7. Are there any questions that seem less important and could be removed from the checklist?
8. Are there any questions that are not part of the checklist, but should be asked when evaluating a score report?

Skills

9. Do you think this evaluation could be completed by one person alone? Who would that person be? What kind of background do they need?
    - What are your thoughts on giving different parts of this checklist to different groups of people, based on their area of expertise?

Other

10. What do you think might be a better format to present these questions?
11. In what contexts do you think this checklist could be used?

C. Score Report Sample