Visibly Learning from Reports: The Validity of Score Reports

John Hattie
Visible Learning Labs
University of Auckland, New Zealand

Abstract

With so many Reports being generated by tests there is a need to consider the validity and principles that maximize the appropriate interpretations that users make from these reports. It is argued that the validity of reports is a function of the users correct and appropriate inferences and/or actions about the test taker's performance based on the scores from the test. The paper presents possible empirical studies to provide such evidence, and introduces a series of 15 principles that aims to assist in maximizing the accuracy and appropriateness of interpretations of Reports.

Over the various editions of the *Standards for Educational and Psychological Testing*, the essence of validity has moved from the characteristics of the test itself, through the interpretation of scores, to the more recent emphasis on the adequacy and appropriateness of inferences and actions based on test scores. No longer do we worry only about whether the "test does the job it was employed to do" (Cureton, 1951, p. 621), but now are more concerned about a more prescriptive set of arguments about whether the decisions based on the test results are defensible (Kane, 2001). The latest set of *Standards* considered that validity referred to the "degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (AERA, APA, & NCME, 1999, p. 9). "The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores made by proposed uses that are evaluated, not the test itself" (p. 9).

These more recent claims have placed more reliance on the user of tests to ensure that they are making appropriate interpretations. We have argued that this view of validity is too restrictive and the pendulum has swung too far from responsibility by the test developer to the test user to provide validity evidence (Maguire, Hattie, & Haig, 1994). Instead, the argument is that the emphasis needs to be placed back on the test developer to provide evidence for the way in which users are not only meant to, but do make evaluative judgments and defensible consequential actions or inferences from the reports based on the scores from tests.

It is the case that the more recent demands of various accountability systems have led to the presence of many more reports – particularly for users who may not be as test sophisticated as could be desired. There are numerous credentialing agencies that provide reports to its users, international test reports, and a growth of testing as entertainment (e.g., "Test the Nation IQ tests", see Cooper, 2003, Fletcher & Hattie, 2010). In schools, Goodman and Hambleton (2004) have noted how the *No Child Left Behind Act* of 2001 has edicted that states must "report" results on mathematics, reading, and science assessments at the state, district, school,

subgroup, and individual student levels across a wide range of grades. Such reports now are being distributed to parents and teachers of about 22 million US students each year. Such reports should be "consistent with relevant, nationally recognized professional and technical standards "(NCLB, 2001, S 1111[b][3][C][iii]) – but there are no such standards for Reports. The International Test Commission has commenced a process for determining such standards, but the process is still in its early stages.

Given this rise in Reports, one of the major differences in the test experience is that the users of these Reports are not necessarily the test administrator, or the person who asked for the test to be administered. The user now can also include the student/user, client, or parent, state accountability office, parents, reporters – thus, throughout the remainder of this paper the notion of "Reader" is used as the generic interpreter of the Reports.

The earliest Reports tended to include many terms, symbols and concepts familiar to test developers but not to teachers, students, or parents. Hence, much energy was placed in finding ways to make meaning of these terms. So often, however, such terms (like standard error, statistical significance) confused, intimidated, or were ignored by readers of these reports (Hambleton & Slater, 1997). Goodman and Hambleton (2004) also noted that many reports were overly dense and reported too much information and many Readers found it difficult to find and extract what was most important and relevant.

This paper outlines a fundamental claim about the validity of Reports, and then via a series of empirical studies introduces a series of principles that aims to assist in maximizing the accuracy and appropriateness of interpretations of Reports. Two other sources of evidence are used to derive and defend additional principals - the human computer interface research and the findings from visual graphics.

Validity of Reports

That there are many readers of reports means that any report must include sufficient information as to the choice of the test, the context of administration of the test, and the psychometric evaluation for the defense of the test scores representing the student's proficiencies presented in the report. In a sense there is a four-part argument operating – that is, there are report validity arguments relating to the (1) choice of the test, (2) the administration of the test (e.g., computer or pencil and paper), (3) the psychometric evaluation of the dependability of the scores, and (4) the accuracy and appropriateness in which the reader interprets this information. All are related, all are important, and all interact with each other. Notwithstanding the importance of the first three parts – these are beyond the scope of this paper and the focus is on the fourth part.

The latest published version of the Standards has little to say about reports. There are admonitions that when "test score information" is released to readers, "those responsible for testing programs should provide appropriate interpretations. The interpretations should describe in simple language what the test covers, what scores mean, the precision of the scores, common misinterpretations of test scores, and how scores will be used" (Standard 5.10). The Standards do stipulate that "when computer prepared interpretations of test result protocols are reported, the sources, rationale, and empirical bias for these interpretations should be available, and their limitations described (Standard 5.11). These statements are more about cautions, they tend to require descriptions and meanings in reports (to protect the report

developer), and they provide little guidance as to the need for or manner of empirical evidence that should be provided. Instead, it is argued that there needs to be much more attention to provide evidence to justify various interpretations *and ensuring that interpretations are appropriately made by the Reader.*

Principle 1: The validity of Reports is a function of the Reader's correct and appropriate inferences and/or actions about the test takers performance based on the scores from the test.

This claim places much reliance on the developer of Reports to provide compelling evidence that Readers make correct and appropriate inferences and actions based on the Reports provided. To address this claim about validity it is minimal to provide evidence that Readers correctly answer two major questions: *What do you see? What would you do next?* These two questions focus on the two most critical aspects of validity: the appropriate interpretations and actions from Reports; and how Readers answer these questions should be shown to align with the intentions of the Report developer. The current trend that asks for more description, more explanation is misleading as the provision of such information may or may not lead to Readers making correct and appropriate inferences – indeed it may lead them to ignore this very information.

There can be many sources of validity evidence to ascertain whether a Report can lead to valid interpretations to these two questions. Three of these methods include tests of the interpretation of Reports, focus groups, and research from the human-computer interface and visual design studies. From these discussions various principles for developing worthwhile and valid Reports are suggested.

The sources of validity evidence relating to Reports

There are few empirical studies relating to how users actually interpret reports. Impara et al. (1991), for example, found that teachers had most difficulties interpreting scale and normal curve equivalent scores, percentiles, and grade equivalent scores (the message was to *minimise* the use of such numbers). Jaeger, Gorney, Johnson, Putnam and Williamson (1994a,b) designed prototype report cards using two formats (tabular and narrative) and two lengths (long and short). The tabular reports made extensive use of tables and graphs and contained little narrative, whereas the narrative reports used few graphics and included mostly descriptive passages. The long reports were 2.5 pages and the short reports were 1.5 pages. The same material was presented in all four types of reports. The reports were also varied by the quality of the report—good, mediocre and poor. Focus groups of parents were asked to read and provide their interpretative comments on these reports. Parents preferred long over short, and narrative over tabular reports—although it was length that was the major consideration (they preferred long over short, and then the narrative reports). The parents, however, most accurately interpreted the longer narrative reports. Similarly, Salvagno and Teglasi (1987) found that teachers preferred interpretive rather than factual information in reports about students. It seems that parents' and teachers' preference and accuracy is related to the format of the reports – provide the interpretation and not merely present the data.

Much of the considerations in the remaining of this paper relate to the development of a national assessment system developed for elementary and high schools in New Zealand. The tool named asTTle (assessment tools for teaching and learning) has been designed to evaluate

and improve the teaching of reading, writing and mathematics in schools through computer based assessment and reporting. It allows teachers to custom build a test from a bank of 10,000 IRT calibrated items using a linear programming heuristic to select items to maximize the fit to a desired test characteristic curve (via their choice of overall difficulty), and with various constraints such as curriculum content and objectives, time, format (adaptive, pencil paper, on screen), surface and deep complexity, open and closed items, minimizing previous item exposure (see Hattie, 2006; Hattie & Brown, 2008; www.asttle.org.nz). asTTle provides various Reports of what students can and cannot do, what teacher have and have not taught well, to whom, and helps set targets for their teaching, shows student growth over time, and suggests resources for moving students' toward the learning targets. The reports are based both on normative comparisons and curricula expectations.

Principle 2: That evidence is needed to demonstrate how Readers are interpreting reports.

This principle is a truism, but evidence is often lacking. As part of the evaluation of asTTle Reports, we devised a series of 35 questions to determine whether Readers could correctly interpret and make appropriate actions from the Reports (alpha = .93; Hattie & Brown, 2008). In one study, 176 teachers and principals were given this test after they first engaged in reading the Reports. The mean score was 20.84 or 60% (sd = 8.44), and this was adjudged not sufficiently satisfactory and thus major modifications were made to the reports to increase the mean to over 90%. The highest relationships with successful interpretation were from teachers with a conception of teaching relating to "assessment is powerful for teaching" (r =.34), whereas there was a negative relation for those who had a conception of teaching as something relating to school accountability (r =-.21). This suggests that the purpose for undertaking assessments can be powerful in the correct interpretations of reports. Those who most correctly interpreted the Reports made more positive (r =.30) and less negative (r =-.16) comments about the value and use of assessment reports.

In another study, about half the teachers and principals (from 60 schools) had no and half had the opportunity to attend professional development about the asTTle application (Ward, Hattie, & Brown, 2003). Attending PD increased familiarity with asTTle, enhanced the spread of information to other persons (students, senior management, board, and parents), gave more confidence in creating reading tests and entering scores, and increased the use of console, group learning pathways, tabular reports, and using the reports for future planning. Those who were not offered professional development claimed to prefer to learn at their own pace, and teach themselves. Clearly those who do not attend professional development find ways to defend their own self-learning; but had lower correct interpretations of the Report.

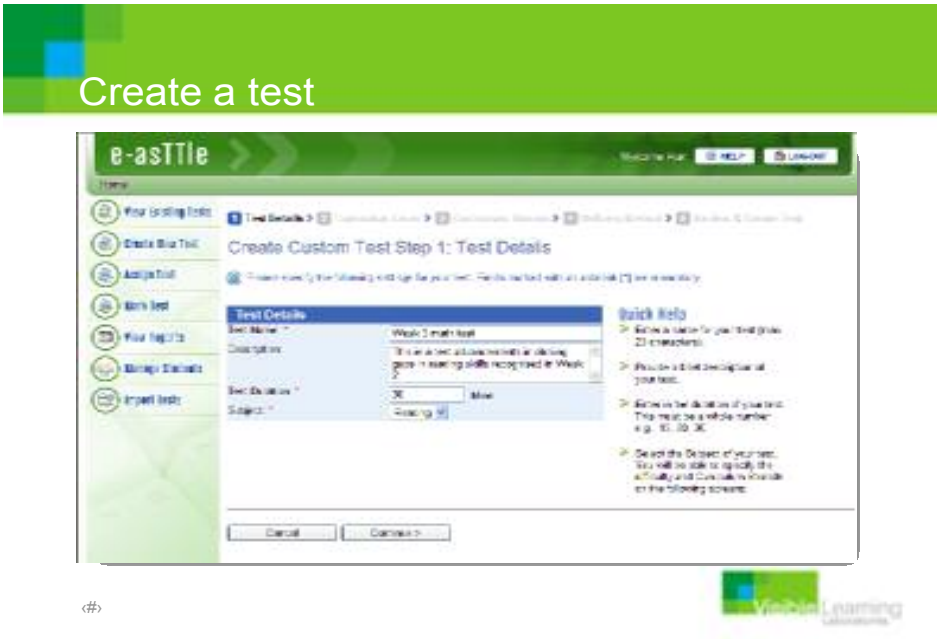
The major form of research in understanding how Readers interpret reports is to conduct focus groups. We used, for example, close to 80 focus groups to devise our own first Reports, but since have tailored these so that much fewer focus groups are needed because of the principles learnt to maximize meaning and actions from Reports. We found that the teachers wanted more guidance on the flow (where to read first, more explanations, clearer titles; Meagher-Lundberg, 2000, 2001) which then highlighted a principle to maximize interpretability of Reports:

Principle 3: Readers of Reports need a guarantee of safe passage.

A guarantee of safe passage implies that Reports need to be created so that the Readers eyes are taken to the most important first, and then to flow to the details. When we asked Readers "What did they see?" we noted their first comments and impressions, and this led to much a greater use of graphics and much greater attention to flow. In our own application we spent a year trialing various forms of flow to ensure a guarantee of safe passage – that is teachers always knew where they were, where to go, where to get out and get in, and whether their interpretations were correct. We asked teachers to go through mock flows, until we found the optimal way to move through the application and reports to maximize the meanings. We ended up with multiple flows – along the top of the screen, down the left-hand guided selection, by checking every forward and back button to ensure proper destination.

Principle 4: Readers of Reports need a guarantee of destination recovery.

Destination recovery allows the user to have confidence in the process leading to the Reports, a sense that they are in control of their use, and that the Reports were generated to provide the information they desired. For example, in the following Report, there are flows down the left hand side, along the top, and the Continue button allows the Reader a guarantee of destination recovery.



‹#›

Principle 5: Maximize interpretations and minimize the use of numbers

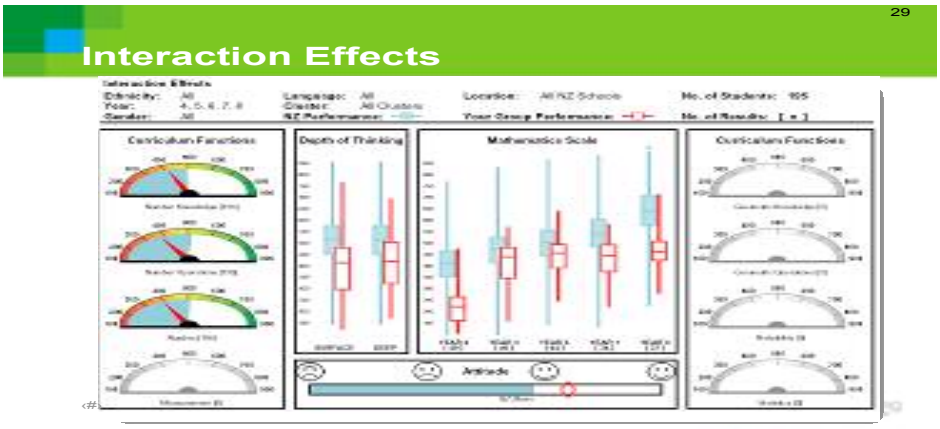
"The greatest value of a picture is when it forces us to notice what we never expected to see" John Tukey (1990).

It is the pictures and words, not the numbers that have the higher probability of leading to valid interpretations. We discovered that the highly graphic nature of reports permitted Readers to access the deeper structure without having to engage in a great deal of linguistic (or numerical to linguistic) decoding. While numbers are excellent summarizes, a downside of the use of graphics was that there needed to be multiple reports, each highlighting different

interpretations and actions. Numbers are meant to be indicators -- and if this is so, then it is recommended that the test developer go straight to the meaning and minimize the use of indicators. If “numbers are words made manifest” then “words are numbers made manifest”. The overuse of numbers in Reports reduces the probability that Readers will make dependable interpretations. Reports full with numbers are rarely of value and they are prone to misinterpretations. Instead, the aim of Reports should be to provide the more appropriate interpretations of these numbers.

Principle 6: Minimize the amount of “numbers” and maximize the amount of interpretations.

The dilemma is that the use of words takes up more space hence the art is providing succinct but dependable meanings of the numbers either via graphics or words. For example, we aimed to include the meanings relating to standard error into the Reports. No matter how we constructed the sentences, conditionals, and trialed these formulations, the users stated that they understood the meaning but they still tended to ignore the error of measurement in their interpretations! Finally, we formulated a pictorial manner for displaying the concept of standard error and found that users made more appropriate interpretations. In the Console Report below, the red arrow indicates the mean for that teacher’s class relative to the “blue” normative sample of similar students. The standard error is depicted by the breadth of the red arrow such that there is a statistically significant difference if they can see white space between their red arrow and the blue norms. The picture rather than the numbers or words told the story: the words were ignored and were indeed not needed to get the correct interpretations.



Principle 7: The answer is never more than 7 plus or minus two

Minimalism as a design option (van der Meij, 1996)

Our working memory has limited capacity, and most of us can process 7 ± 2 bits of information at any one instance. Miller (1956) who pioneered this finding claimed that “my problem is that I have been persecuted by an integer. For seven years this number has followed me around, has introduced in my most private data, and has assaulted me from the pages of our most public journals. This number assumes a variety of disguises, being sometimes a little larger and sometimes a little smaller than usual, but never changing so much as to be unrecognizable (p. 81). There is the inherent “tension between stimulating the user to think about tasks so that he

or she remembers it afterwards (learning), and helping the user complete tasks quickly and without considerable cognitive effect (doing)" (Miller, 1956, p. 19). Our experience is that Reports need to focus on one major theme to maximize the correct interpretations, and within a Report have no more than 7 ± 2 chunks of information.

Minimalism can be articulated in four key principles: choosing an action-oriented approach (e.g., a demonstration that the interpretations of the Reports lead to consequences), anchoring the tool in the task domain (e.g., ensuring it is curriculum not measurement related), supporting error recognition and recovery (see above), and supporting reading to learn and do (e.g., have links to deeper meanings, if desired; Carroll, 1990, 1998). Choosing an action-oriented approach requires a balance between the Report reader's need for knowledge and need for action and consequences – it is the balance of the two key questions – What do you see? and What do you do?. Most teachers are keen to evaluate, act and then do something meaningful, but any interpretation requires knowledge and understanding.

Principle 8: Each report needs to have a major theme

Mixing up major themes leads to Readers missing major inferences, make too many inappropriate connections across themes, and ignore critical conditionals on the overall interpretations of each theme. For example, the Console Report (above) allows for up to eight sub dimensions of a curriculum area. If there were any more then the information overwhelms the Reader, and we observed that when more is provided teachers tend to print the Reports, store them in the folders, and place them on bookshelves. There are few consequential actions of importance to their teaching and learning.

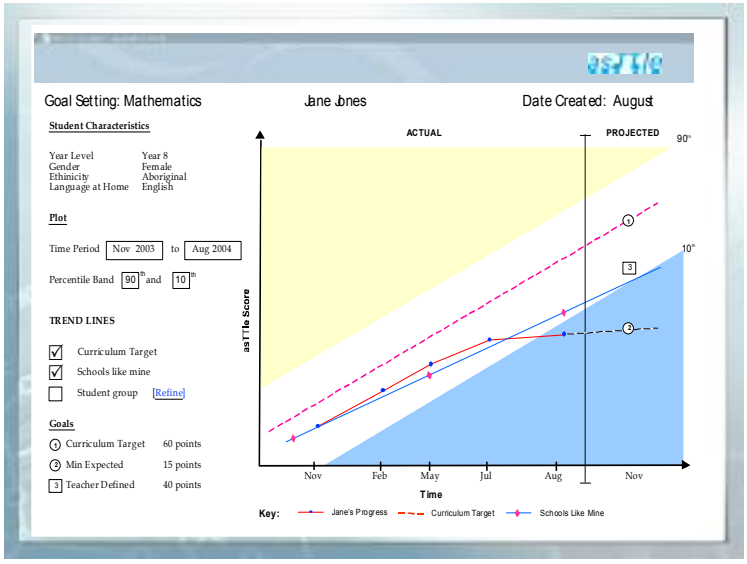
If details are desired, drilling down can occur. Such drilling down can provide more details of a second-order nature. For example, the Console report allowed the Reader to make decisions about which group(s) they wished to make comparative interpretations (e.g., comparing boys with norms about boys or with girls, ethnicity, English as a second language, schools like mine). As another example, we had pressure to provide details about each of the 100+ objectives of the Reading or Mathematics curriculum, but such detail overwhelmed the reader and there was a tendency to gloss over these details and make low level rather than more appropriate higher level interpretations. Instead, we constructed the Individual Learning Pathway in which the many objectives are presented, but the higher order inference is the four quadrants of information – what is this students strengths, weaknesses, achievements, and to be achieved objectives or standards. If desired, Readers can drill down to specific objectives, their meaning, example items, and student performance on the objective. The decision was to NOT allow too much attention to the item level – while the quality of items is critical, and over emphasis on interpreting individual items cannot be sustained given the reliability and low levels of information at an item level. The aim should not be for teachers to re-teach items, but to accumulate information across items at a slightly higher order (e.g., objective, curriculum sub domain) and make inferences and decisions at this level.

It is optimal to have more reports (provided there are no more than 8 at the highest level). We ended up with seven reports at the highest level – each highlighting a major theme for Readers to focus on and interpret:

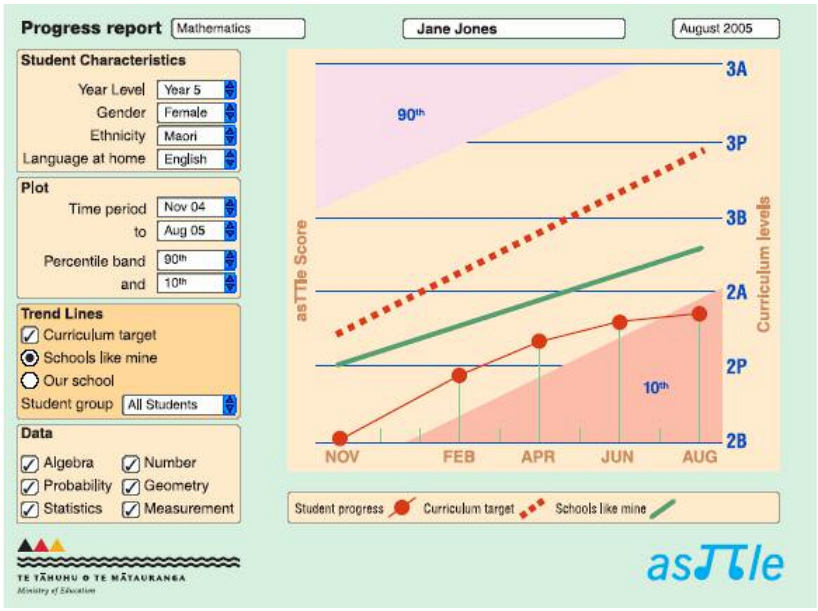
Console Reports	How am I going relative to others?

Individual or Group Learning Pathway	How am going relative to my strengths, weaknesses, achieved and not achieved?
Curriculum Skyline	How is this group distributed in their achievement?
Progress	How am I going over time?
Expectation	What are the student/teacher targets over time
Tabular Report	Downloadable for further analyses and storage.
What Next	What do I do next?

There can be a temptation to provide pretty Reports, fill of meaningful information, and each part of the Report sensible. This can be an error *if* the Reader does then not get the correct higher level inferences. As an example of the necessity to reduce the amount of information to maximize the correct derivation of the key inferences, consider the following report about student progress. This Report was very appealing and liked by the teachers, but the teachers spent too much focus on how they could manipulate the information (down the left hand side) and tended to not be as concerned about the more critical decisions within the actual graph. Thus, we needed to remove a lot of the functionality on the left, and spend more time concerned with the meanings on the right.



Much of the left was redesigned in thumbnails (to allow access to more information after the major theme was obtained), greater colour contrast, and increased headings. Even this was too much and we failed the first test “What do you see?”



This above Report was still too cluttered and lead to Readers to want to drill down immediately and thus make more refined interpretations before they inferred the overall picture (of this students slowing down in their growth in the last 3 months). The final Report was much simpler:

Principle 9: Anchor the tool in the task domain

It is argued that the task domain for teachers when they interpret reports is "teaching and learning" and not "assessment". Hence it is argued that there is no need for "assessment literacy" as teachers need not be required to learn the language of psychometricians. Instead test report developers need to learn the language of teachers, which is teaching and learning. The claim is that it is not the assessment literacy of the teacher, but the use by the test developer to use the literacy of the Reader.

Design principles of reports to maximize valid interpretations of Reports.
This section outlines a series of principles derived from research on the human-computer interface and visual literature to maximize the design of Reports to thence facilitate appropriate interpretations.

Principle 10: A Report should minimize scrolling, be uncluttered, and maximize the "seen' over the 'read'.

Leeson (2006) reviewed much of the human computer interface (HCI) literature and found that there were some factors that assisted or impeded the readability and interpretability of Reports from computers or computer generated Reports. The key factors were scrolling, which had a detrimental impact on performance (and was unrelated to general ability), and the lack of facility to review or underline parts of the screen, which did not affect performance but did affect the confidence users had in the verisimilitude of the Reports. It is also found that users are more likely to *scan* information on screens whereas they *read each line* on paper, and thus there is a need more whitespace and clear and crisp graphics (at least 2pt interline spacing, use of one-column blocks), larger font size (adults perform best on Times New Roman, Arial, Tahoma at 12 point, and children on Arial and Comic at 14 point size), higher screen resolution (and this also reduces fatigue), and presenting one item per screen with cues to other or related items is critical. Reading speed of the information on screens was optimized for adults when line length falls between 75 and 100 characters per line (cpl), and for children at 45 cpl.

The literature on effective visual displays reinforces these claims about the need to highlight the important by minimizing the clutter (e.g., Wainer, 1997; Tufte, 1983, 1990). Principles include: keeping the report uncluttered, using text to support and improve the interpretations of charts and graphs, minimizing statistical jargon, including a glossary of key terms, using bar charts to facilitate comparisons, grouping data in meaningful ways, using boxes or graphics to highlight main findings, avoid the use of decimals, use color in a purposeful manner.

Wainer, Hambleton, and Meara (1999) redesigned graphic data displays from the 1994 National Assessment of Educational Progress (NAEP) reading assessment to improve their communicability. In a controlled experiment, it was found that visual displays redesigned according to Wainer's (1997) principles yielded more accurate and faster answers. The principles advocated include: (a) displays should allow access to the user's deep structure without recourse to linguistic decoding – that is, they should be "seen" and not "read"; (b) displays should have a clear communicative purpose; (c) displays should not attempt to do too much; and (d) displays should avoid clutter (e.g., footnotes, notations). Respondents indicated a preference for bars rather than lines and for the use of color. In addition, visual displays that featured "bigger print, additional white space, [and a] "lack of … footnote[s]" (Wainer et al., 1999, p. 319) were also preferred. Accumulated displays and footnotes were seen as

inappropriate despite the fact that they may display information in a more accurate way. Finally, it was concluded that conventional displays with which users were most familiar tended to be preferred

Based on analyses of many Reports, Goodman and Hambleton (2004) outlined a number of features that they considered made reports more readable such as: use of headings and other devises to organize reports, using highlighted sections, use of graphics, and personalizing the reports. Problems included: reporting too much information, use of statistical jargon, lack of information regarding the purpose of the assessment, how test results will be used, and precision of test scores.

Principle 11: A Report should be designed to address specific questions.

It should be clear about what are the questions that the Report aims to illuminate. Each Report should have a clear set of potential interpretations desired by the developer of the Report Reports, Goodman and Hambleton (2004) suggested:

a. Given this report, what would you do next? – and this should align with intended consequences, make the report has more than "interesting" status

b. How would you navigate around the Report screen? – something about options helping clarify, not cluttering, not allowing message to be diminished (e.g., using comparison to make one look better just for this reason)

c. What potential uses can you see for this Report? The uses are coded for validity relative to the quality of the data in them

The National Education Governance Group (1998, p. 31) recommended:

d. How did my child do? What types of skills or knowledge does his or her performance reflect? How did my child perform in comparison to other students in the school, etc.? What can I do to help my child improve?

Hattie (Hattie & Timperley, 2008) argued that there were three critical questions underlying accomplished teaching, learning, and assessment and they can be used also to guide the development and the validity of the interpretations of Reports:

e. Where am I going? How am I going? Where to next?

Principle 12: A Report should provide justification of the test for the specific applied purpose and for the utility of the test in the applied setting.

Any test user should have an expectation that a report about their performance would be made available in a timely manner. In some cases the Report could be made available to the person who requested the assessment (in today's computerized delivery the test administrator may not "exist"), and in other cases also to the person who is tested.

Principle 13: A Report should be timely to the decisions being made (formative, diagnostic, summative and ascriptive).

Principle 14: Those receiving Reports need information about the meaning and constraints of any report.

Depending on who is to receive the Reports, then there needs to be appropriate training and assistance to ensure that the meaning and constraints of any Report are understood by the Reader of the Reports.

Principle 15: Reports need to be conceived as actions not as screens to print.

The above claims are that the validity of Reports relates both to the Reader making the correct interpretations and making appropriate consequential decisions. Thus, the orientation is 'doing' vs. 'printing'. There is also a need to differentiate between a learning effect (when the Reader first confronts a Report, and a re-use effect, when they have learnt the correct higher order interpretations and then the Report should lead to more appropriate more refined interpretations, to the degree possible). A key finding from how Readers move from first to later uses is that the details and information about the details need to be readily available to the Reader – just in time. For example, Readers have a 40% reduction of training time and 50% more basic tasks completed successfully from a minimal manual and HELP than from the use of a Manual (Carroll, Smith-Kerker, Ford, & Mazur-Rimetz, 1987; van der Meij, 2008). There is a high level of success from a personal style of the "dummies" approach – speak to the Reader as if another is speaking to you. This is the "co-user" approach where this Help is a plausible colleague.

Conclusions

Ryan (2009) commented that score reports need evidence to support evidence of the interpretations, they can be created and improved with systematic research, they require "specialized" language, and most important validity is an attribute of the inferences and the uses of assessment information relative to the purpose of the assessment. The reports are the vehicle for transmitting assessment information to those making the inferences and using the results. Assessment validity is improved by developing or modifying the reports to increase the likelihood of appropriate inferences and proper uses of assessment information. As noted in this paper, there needs to be emphasis on both the descriptions made from the reports, and on the literal comprehension that are inferred. There needs to be evidence to defend the reasonableness of inferences and appropriateness of intended actions.

The claim is that the validity of Reports is fundamentally related to a preponderance of evidence relating to the accurate and appropriate interpretations that users make from assessment reports. This evidence is not related so much to the quality (utility, accuracy, feasibility, propriety) of what is presented but on the actual interpretations that users make from the Reports. An emphasis on what is presented may lead to misuse (as in the example of providing definitions of standard error, noted above), clutter, and limited usefulness of Reports. The current trend asking for more description and more explanation is misleading as the provision of such information may or may not lead to Readers to make correct and appropriate inferences – indeed it may lead them to ignore this very information. There are many methods that can used

for this purpose; such as empirical research studies, focus groups, case studies, content analysis, interviews, and developing assessments of the interpretations.

The fundamental claim is that the validity of Reports is a function of the Reader's correct and appropriate inferences and/or actions about the test takers performance based on the scores from the test. The primary inference and action relate to "What do you see?" and "What would you do next?" A series of research methods is presented and other similar methods are encouraged. Such methods include empirical tests of the interpretation of Reports, and focus groups to develop and evaluate the inferences from Report. A series of 15 Principles for reports are recommended relating to design principles of reports aiming to maximize the valid interpretations of Reports.

References

AERA, APA, & NCME (1999). *Standards for educational and psychological testing*. Washington, D.C.: Author.

Carroll, J.M. (1990). *The Nurnberg Funnel*. Cambridge, MA: MIT Press.

Carroll, J.M. (1998). Reconstructing minimalism. In J.M. Carroll (Ed.), *Minimalism beyond the Nurnberg Funnel*, (pp.1-17). Cambridge, MA: MIT Press.

Carroll, J.M., Smith-Kerker, P.L., Ford, J.R. & Mazur-Rimetz, S.A. (1987). The minimal manual. *Human-Computer interaction, 3*, 123 - 153.

Clarke, S., Timperley, H., & Hattie, J.A. (2003). *Assessing formative assessment*. Hodder Moa Beckett, New Zealand.

Cooper, C. (2003). *Test the Nation: The IQ Book* London: BBC Books.

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621–694). Washington, DC: American Council on Education.

Fletcher, R.B., & Hattie, J.A. (2009). *On Intelligence*.

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

Hambleton, R. K., & Slater, S. C. (1997). Reliability of credentialing examinations and the impact of scoring models and standard-setting policies. *Applied Measurement in Education, 10*, 19-39.

Hattie, J.A.C. (2006, July). *Large-scale assessment of student competencies*. Paper presented at part of the Working in Today's World of Testing and Measurement: Required Knowledge and Skills (Joint ITC/CPTA Symposium). 26th International Congress of Applied Psychology, Athens, Greece.

Hattie, J. A. C. & Brown, G. T. L. (2008). Technology for school-based assessment and assessment for learning: Development principles from New Zealand. *Journal of Educational Technology Systems, 36(2),* 189-201.

Hattie, J.A.C., & Timperley, H (2008). The power of feedback. *Review of Educational Research*, 77(1), 81-112.

Impara, J. C.; Divine, K. P.; Bruce, F. A.; Liverman, M. R. & Gay, A. (1991). Teachers' ability to interpret standardized test scores. *Educational Measurement: Issues and Practice, 10 (4),* 16-18.

Jaeger, R. M., Gorney, B., Johnson, R., Putnam, S. E., & Williamson, G. (1994a). *A Consumer Report on School Report Cards.* The Evaluation Center, Western Michigan University.

Jaeger, R. M., Gorney, B., Johnson, R., Putnam, S. E., & Williamson, G. (1994b). *Designing and Developing Effective School Report Cards: A Research Synthesis.* The Evaluation Center, Western Michigan University.

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

Leeson H.V. (2006). The mode effect: A Literature Review of Human and Technological Issues in Computerized Testing. *International Journal of Testing, 6(1),* 1-24

Maguire, T.O.M. & Hattie, J.A., & Haig, B. (1994). Construct validity and achievement assessment. *Alberta Journal of Educational Research*, *40*, 109-126.

Meagher-Lundberg, P. (2000). *Comparison variables useful to teachers in analysing assessment results.* (Tech. Rep. No. 1). Auckland, NZ: University of Auckland, Project asTTle.

Meagher-Lundberg, P. (2001). *Output reporting design: Focus group 2.* (Tech. Rep. No. 10). Auckland, NZ: University of Auckland, Project asTTle.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81-97.

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. (2009, April). *Making test score scales and reports: More understandable and useful*. Discussant presentation at the National Council on Measurement in Education, San Diego, California.

Salvagno, M. & Teglasi, H. (1987). Teacher perceptions of different types of information in psychological reports. *Journal of School Psychology, 25*, 4l5-424.

Tufte, E.R. (1983). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.

Tufte, E.R. (1990). *Envisioning information*. Cheshire, CT: Graphics Press.

Tukey, J.W. (1990). Data-based graphics: Visual display in the decades to come. *Statistical Science, 5(3), 327*-339.

Van der Meij, H. (1996). Does the manual help? An examination of the problem solving support offered by manuals. *IEEE Transactions on Professional Communication*, *39*(3), 146-156.

Wainer, H. (1997). *Visual revelations: Graphical tales of fate and deception from Napoleon Bonaparte to Ross Perot*. New York: Copernicus Books.

Wainer, H., Hambleton, R.K., & Meara, K. (1999). Alternative displays for communicating NAEP results: A redesign and validity study. *Journal of Educational Measurement, 36(4),* 301-335.

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/Ministry of Education.