

**Score Reporting, Off-the-Shelf Assessments and NCLB: Truly an Unholy Trinity**

**By**

**Jon S. Twing, Ph.D.**

**Pearson**

**Paper Presented at the Annual Meeting of the National Council on Measurement in  
Education (NCME), March 25, 2008, New York, New York.**

### **Abstract**

One consequence resulting from NCLB, particularly as instructional time becomes more precious, is the desire to be more efficient in assessing learning. Augmented shelf-assessments would seem to be a logical choice in reaching this efficiency. This paper outlines some potential pitfalls in this thinking as it applies to making valid inferences from scores resulting from augmented shelf-assessments. The risk regarding these potential pitfalls sometimes enhanced by a desire to squeeze more and more information from the assessment results themselves, regardless of the purpose originally intended for the measures. The current paper provides relevant research and discusses issues affecting the reporting of scores, including: obtaining diagnostic information from assessment results; linking assessment results to instruction; and ways to improve reliability of part-score interpretations. The paper concludes by presenting common teacher expectations regarding the use of the results and suggesting that more research is needed regarding score reporting for augmented shelf-tests.

## Introduction

If one were to look closely they would see that small and subtle, but perhaps powerful positions are being articulated by the United States Department of Education (USDOE) and others regarding the future of assessment, particularly assessment for accountability. NCLB, while not dead yet, is likely to morph (or has already spun its cocoon of change) into something very different. For example recent rhetoric from the USDOE shows a substitution of the traditional and familiar phrase of “all students proficient by 2014” with the more difficult to achieve slogan of “on grade level by 2014” (Holly Kuzmich, Personal Communication, 2008). Other topics like college readiness, 21<sup>st</sup> century skills, end-of-course or end-of-instruction examinations, and growth modeling have all but supplanted the attention placed on the basic requirements of NCLB.

Given what seems to be a dying issue, why then discuss anything related to NCLB (regardless if it is specific to shelf tests, augmented norm-referenced tests (aNRT) or customized standards-referenced (SRT) assessments)? An anonymous reviewer of the proposal for this paper caught the essence of the dilemma and why NCLB is still important (and specifically why results reporting are important) with a very simple phrase: “Over-reporting underspecified constructs remains a problem in the request for more ‘diagnostic’ feedback.”

After all, the reason one might choose an augmented NRTs is to add more measures of the constructs (standards and benchmarks) to an existing “shelf-test” such that it better aligns with state curriculum. The idea being that we get to have our cake and eat it too...by keeping the efficiency and presumed economies of scale from using an existing shelf-test (in this case one with norms) and by adding items so that it can be a perfectly aligned measure of our selected content standards. Naturally, any reporting done regarding the performance of individuals on such an augmented assessment would be presumably better for at least two reasons, norms and feedback regarding how well students performed on perfectly aligned and hence presumably instructed content. Implicit in the argument is the assumption that the norm-referenced scores resulting from an augmented NRT are valid for the purposes intended and that the combined scores are a valid measure of the content standards and benchmarks in a traditional standards-referenced sense. Traditionally, when confronted with this need in the past, researchers have rallied around the position that if the rules of standardization within an administration remain intact, then the norms associated with that section carry little or no threat to the validity of their

interpretation (Linn & Hambleton, 1992; Yen, et. al., 1987). In other words, the use of intact “testlets” of NRT blocks under restricted circumstances and context are likely to yield valid norm-referenced scores if the circumstances are controlled properly. If the Mathematics Reasoning section of a shelf-assessment, for example, has a raw score to percentile rank table and was administered in a self-contained section following the rules of standardization, then administering it alone or intact within another assessment should not pose much of a threat to the resulting score interpretations. This is fine for the NRT piece of the cake, but what about the other pieces or the entire cake itself? Will the use of essentially three sets of scores, one from the NRT section, one from the augmented section and one based on both sections combined result in an over-reporting of the under represented domain of mathematical reasoning? Other research suggests that the degree of content match between the augmented or customized assessment and state standards have varying impact on the resulting score interpretations (Forsyth, Twing & Ansley, 1992; Yen, Green, & Burket, 1987; Linn & Hambleton, 1992).

The position taken by this paper suggests it might not be the case that an augmented assessment, comprised of a shelf assessment plus some additional items, is actually a better reflection of the content domain being measured regardless of the accuracy of the score interpretations resulting from the shelf piece. The compromises required to construct such an assessment might actually lead to a change in the construct definition of the desired measure and resulting inferences of the scores from such a measure might be in error. Considerations that extend results interpretation to the sub-domain level are reviewed in light of this argument.

### **Reporting Perspectives Required for Valid Inferences**

Arguably, it is fair to say that the current joint technical standards (AERA, APA & NCME, 1999) consider the reporting of results from an assessment as the essence of validity. Consider, for example, the very first paragraph in the chapter on validity from the Standards (AERA, APA & NCME, 1999, pg. 9):

“Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score

interpretations. The proposed interpretation refers to the construct or concepts the test is intended to measure (*emphasis added*).”

At face value one is struck with the importance of reported test scores, not as indicators, metrics or quantifications, but rather as linked to their resultant interpretations. Clearly, the intended or proposed score interpretations referenced by the standards are completely dependent upon the construct being measured. Hence, the definition of this construct (as operationally defined as the domain of skills, attributes or abilities being referenced) and the sampling of items or tasks from the construct, is equally important and a pre-requisite for inference. This point is supported from another section of the Standards in the same validity chapter (AERA, APA & NCME, 1999, pg. 10):

“Construct under representation refers to the degree to which a test fails to capture important aspects of the construct. It implies a narrowed meaning of test scores because the test does not adequately sample some types of content, engage some psychological processes, or elicit some ways of responding that are encompassed by the intended construct (*emphasis added*).”

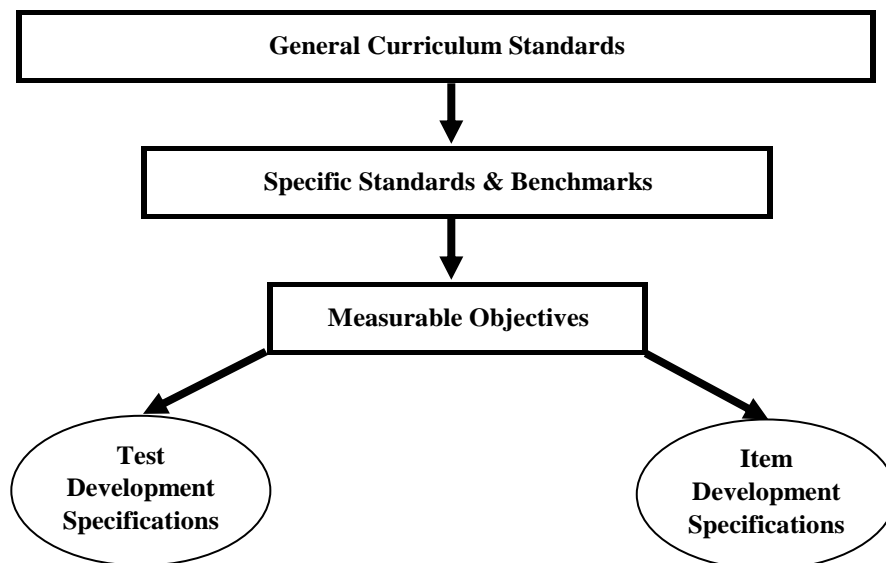
The Standards provide an example from reading ((AERA, APA & NCME, 1999, pg. 9). Poor representation of the intended construct to be measured, in this case reading ability, might be due to the lack of the adequate sampling of reading stimuli (i.e., one inference might be made if the reading passage is a narrative while another and different interpretation might be made if the reading passage is technical). Such a mismatch with the intended construct essentially causes incorrect inferences to result—but not necessarily incorrect scores. The student might have answered all items associated with the narrative passage correctly, with one level of reading skill inferred whereas had a more technical passage been sampled a different score and a different inference would likely result. As such, adequate sampling of the construct or intended domain or skill set to be measured is paramount to the valid interpretations of resulting scores.

#### Domain Sampling

Consider Figure 1 which depicts a typical funneling or reduction of curriculum standards into an assessment design. Traditionally, large scale efforts toward measurement start with an explicitly stated curriculum. Similarly, shelf-tests like NRTs, also link to a set of content standards or instructional objectives. In the specific case of NRTs this can be described as some version of a “consensus national curriculum”. Under NCLB, statewide curriculum is the focus

with some states having well developed curricula (Texas, Florida, California for example), which may have gone under some revision due to the requirements of NCLB, but otherwise which have been in place for a long time (see for example Cruse & Twing, 2000). Other states have evolving curricula (Mississippi, Minnesota) with revisions or multiple revisions having recently been or are currently being implemented. Iowa has only recently addressed the debate about state mandated curriculum or content standards (Iowa Senate File 2216, 2008)—albeit by virtue of the fact that somewhere in excess of 95 percent of all schools in Iowa use the Iowa Test of Basic Skills implying a default curriculum of sorts. Regardless of the stated curriculum or instructional goals, seldom is it likely that the enormity of the entire curriculum can be taught, let alone assessed. Hence, designing assessments (as well as instructional units) will probably undergo a reduction process. Figure 1 presents what might be a logical reduction path in this regard.

**Figure 1. Example Reductive Measurement Path.**

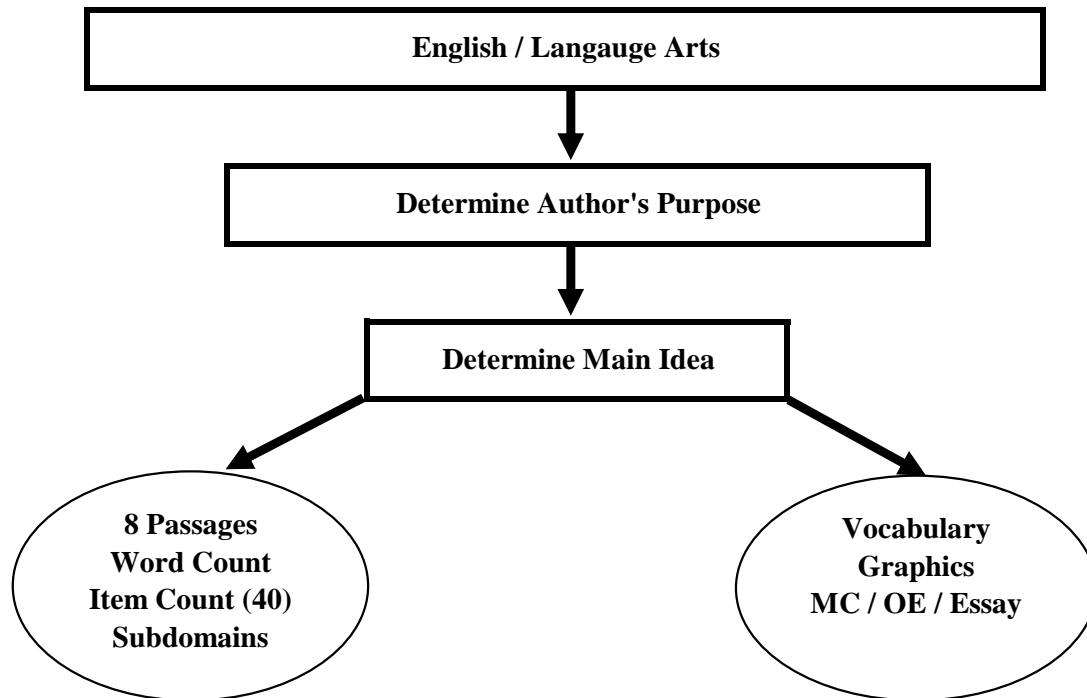


Consider a contrived example extending this concept. Start with the overall goal of instructing and measuring say, English/Language Arts (ELA). Perhaps the general curriculum explicitly states ELA, including reading and writing, as the overarching construct to be learned and tested. One operational definition of what comprises at least one and possibly more

components of ELA might be the still general domain of “Determine, argue and support with evidence author’s purpose or point of view when presented with author generated text”. From this teachers might understand better the instructional objectives (read for understanding, take a position, support a position with evidence from the passage, present arguments for the position, etc.). From an assessment perspective, decisions are still required. For example, we can perhaps construct multiple-choice test questions to see if the student took a position regarding the author’s point of view. Arguably, however, we would need an essay assessment or a portfolio of work in order to evaluate the strength/weight of the evidence and or the strength of the argument. Hence, another reduction to more the more specific requirements of an assessment or evaluation system is needed. These could be termed “test development” or “item development” specifications. Such specifications would delineate more specific information regarding the construction of the test questions (from the domain of skills) and the test form itself. Such things as the names and descriptions of specific content objectives like “Main Idea”, number and types of items, number of literature selections and type of selections, etc., would all be included in this further reduction/articulation of the content domain (construct) into an operational assessment. This example process is provided in Figure 2.

Figure 2 presents, in one oversimplified flow, how we can move from a grand conceptualization of the construct (English/Language Arts in this case) to a very specific example of a 40 item test. In this example, we have moved from literature, poetry, functional reading, writing, graphical organizers, collaboration, spelling, grammar, vocabulary, reading comprehension, fluency and all the other things that comprise ELA to a much smaller sampling of expected student behavior confined within the constraints—albeit self determined constraints, but constraints nonetheless—of large-scale assessment. Is it no wonder that the scores resulting from such an assessment and more importantly, the inferences made from such scores, are likely to be questioned regarding their importance, meaning and ability to convey the learning status for an individual?

**Figure 2. Example ELA Reductive Measurement Path.**

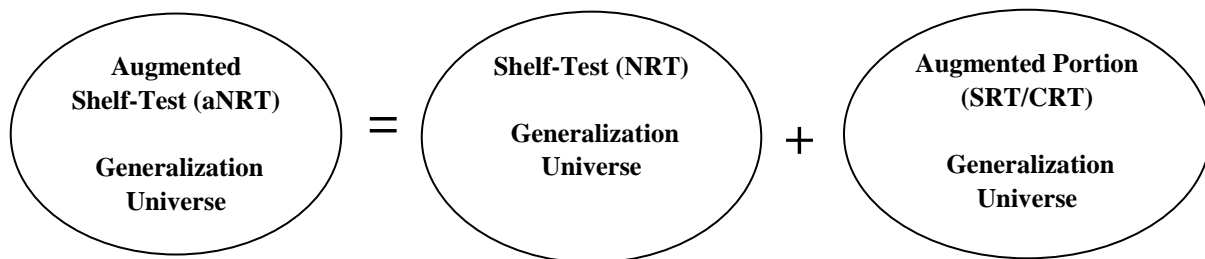


#### Inference Generalization

As if this process of operationally defining the construct in terms of measurable skills amenable to testing was not difficult enough, the concept of commingling a shelf test with another “standards-referenced assessment” adds even more complication. This process is very cumbersome operationally. For example, start with a nationally standardized norm-referenced assessment (this could also be the ACT or the SAT or any shelf assessment). Presumably this assessment measures some construct or set of constructs articulated via a stated curriculum, set of content standards or benchmarks or other instructional objectives. Presumably, since this is a shelf product it is applicable to multiple venues. For NRTs what is being measured is typically characterized as a national consensus curriculum. This curriculum may be attempting to define ELA in terms of commonalities observed across specific state curricula for example. As such, these content standards will match or will be aligned to some extent with a given state’s curriculum, content standards and benchmarks. Let’s assume an alignment study was done, agreement was obtained, and that in general there is 85 percent overlap (alignment) between the

NRT and this example state’s content standards. Under NCLB’s requirement of perfect alignment, this NRT could not be used to measure the state content standards alone and hence would need augmentation. On the other hand, if the state decided to construct their own standards-referenced test (SRT) from scratch (requiring item development, field testing, research, scaling and equating, standard setting, essentially all the steps to make a legally defensible assessment; see for example Smisko, Twing & Denny, 2000) they would ensure complete alignment but presumably at a greater cost. So, driven by the desire to be completely aligned yet benefit from the economies of a shelf-test, this state might choose, instead to develop an augmented NRT (aNRT). Under this model the state would construct from scratch items that would “fill the gap” of 15 percent missing content coverage left out of the NRT. While the logic of such an endeavor is self-evident, considerations regarding the domain of greater content being sampled must be discussed. This concept is represented as articulated in Figure 3.

**Figure 3. Generalization Universes for Augmented Shelf Assessments.**



The untested premise depicted in Figure 3 is that the domain sampled from the augmented shelf-test is the same as the union of the separate shelf and augmented domains. There are some circumstantial reasons to believe this might not be the case as well as some research suggesting it is indeed not the case. The circumstantial evidence is the need to make wholesale compromises in the format of the testing situation when including shelf, particularly NRT components. To match standardization criteria perfectly, things such as page layout, timing, instructions, item formats, white space, answer documents, etc., need to remain constant on the NRT in order to protect the validity of not only the norm-referenced scores, but the also the artifacts of the items themselves (item parameters, classical statistics, equating and scaling methods) which could impact the total score. It is highly unlikely that the same set of specifications would result had

the state constructed an SRT or CRT from scratch (i.e., that the domain sampling would be the same). Even if the augmented and NRT portions are in different sections it is doubtful they will appear the same or will have the flow of an otherwise constructed from scratch SRT. Furthermore, issues of sequencing and context will also play an unknown role in generating a total or part score from the pressing together of two disparate assessments.

Research indicates (albeit dated research) that selecting locally developed items and/or items from an NRT that match specific curriculum frameworks have an effect on the resulting interpretation of percentile ranks (see for example Forsyth, Twing & Ansley, 1992; Twing, Forsyth & Ansley, 1990; Way, Forsyth & Ansley, 1989; Way, 1988; Harris, 1987). Often the difference in customized (augmented or content-specific selected items) and NRT percentile ranks were greater than 10 percentile rank points, though this impact was not universally positive or negative (i.e., scores both went up and down).

Setting aside the percentile rank resulting from the NRT portion, how will the total, sub-domain or part-score (i.e., the standards-referenced piece combining the aligned NRT items and the augmented additional items). It would seem too convenient to dismiss the impacts documented in these studies as related differences in learning of the population of test takers (i.e., the standardization population vs. the more specifically trained local or statewide population) only. Student's scores, raw scores as well as percentile ranks, were different depending upon the customization performed. Hence, it would appear to matter how the customization or, in our extended example, the augmentation of the NRT is to be performed and the validity of these augmented scores also comes into question.

### **Score Reporting**

The previous section attempted to establish the validity of scores as part-and-parcel the definition of the construct being measured and the sampling of the refined content domain operationally implemented in an assessment. The argument was made that inferences from a combination of domain samplings would not necessarily result in similar inferences to one sampling from the more generalized domain. This argument was based in two explicit parts, one circumstantial, one from research as well as a logical implicit argument that this could be the case. This section focuses on the need for scores linked to assessment and discusses the validity of score inferences in this regard.

### Historical Perspective

Aside from the discussions raised in the previous sections regarding domain sampling and inference generalization, what can be learned from an historical perspective regarding score reporting? First, we find that what was old is new again. Consider the following from the 1950 edition (first edition) of Educational Measurement (Lindquist 1950) from the chapter by Walter Cook (Cook, 1950, pg 28) with contributions by William McCall and Ralph Tyler:

“The testing program of the school should be systematic and comprehensive. It should furnish the teacher with up-to-date information regarding the growth record and status of each of his pupils...The tests should measure at regular intervals the permanent learnings which have been achieved...(emphasis added).”

Three major themes appear as early as 1950, namely, the timeline for the reporting of results, the measurement of growth in addition to status; and the need for measurement to occur at regular intervals—implying a link to instructional units or learning. The recent call for more formative assessment would suggest that such information is still desired (and hence has not been routinely produced, even under NCLB (Education Week, 2006).

Cook’s chapter (Cook, 1950, pg 28) continues with more sage advice:

“The results of the testing should always be portrayed in graphic profile form, showing the growth in each differentiated ability from year to year....Systematic testing of the type described above should preferably be done at the beginning of the school year...(emphasis added).”

The graphic reporting of results, even the very simple use of histograms or pie charts is still rare in the reporting of most large scale assessment results. When it is reported it is typically relegated to a technical manual. Few large scale assessment programs test in the fall of the year, let alone get results back in time for a teacher to use during instruction.

Cook and his colleagues did not forget about formative or diagnostic assessment either (Cook, 1950, pg 29):

“Other tests of a more diagnostic nature...should be available to teachers at all times in order to determine individual needs and measure progress in the skills and abilities being emphasized in instruction. These tests should always be selected and used strictly from the standpoint of their instructional value. (emphasis added).”

## Running Head: Results Reporting

From this selection, we can see that almost 60 years ago diagnostic tools were needed for teachers in order to tailor instruction and measure growth. The perception of the value of such assessments was relative to instruction and not status or accountability. Assessments under NCLB struggle to fulfill even partial aspects of what Cook and others claim “should” be the goal of assessment.

Cook also had some recommendations regarding the types of scores reported for individual students (Cook, 1950, pg 30):

“The plan of reporting to parents percentage marks, or letter grades is not consistent with the policy of meeting the needs of individual pupils. Likewise, the practice of simply marking a pupil satisfactory or unsatisfactory in terms of his general learning capacity is inadequate (*emphasis added*).”

If we were to substitute “proficient” for satisfactory in this selection it could apply directly to the current practice under NCLB. Does this not imply that NCLB assessment then, by design, will not meet with the policy needs regarding the individual student? The rhetoric of NCLB is that “no child will be left behind” while in reality this legislation seems to be intended for school improvement or program accountability (Henry Johnson, 2008, Personal Communication). As such, the argument could be extended that any reporting under NCLB is not intended to improve or meet the needs of the individual students but is really for accountability purposes or program improvements. This is not inconsistent with the goals of the legislation, but the rhetoric seems to be about all children learning as much if not more than program accountability.

Cook specifically address the issue of reporting scores for students from part tests or assessment batteries (Cook, 1950, pp. 36-37) which might be applicable to the context of this paper regarding shelf-tests or augmented assessments:

“...such achievement test batteries are too general to be used as a basis for instruction even when detailed analysis of items is made...the sampling of items is too limited and the organization too gross for such tests to be considered as adequate guides in the planning and direction of educational experiences for individual pupils. (*emphasis added*).”

This selection, in a nut shell, foretold the need and desire for augmented achievement test batteries. In other words, if what Cook says is true, additional items must be included with the assessments in order to be more specific (i.e., match the content standards better) and to have

sufficient sampling of the curriculum (as well as student behavior) in order to make statements (inferences) about individual student performance.

#### Sub-score Reliability

Sub-domain scores are desired by teachers, parents and students in evaluating strengths and weaknesses (Yen, et al., 1997; Wainer et al., 2000). Regardless of the type of assessment, there are clearly concerns regarding the basic measurement properties of reliability (in addition to score inference or validity) for any set of sub-domain, part-score or objective-level interpretations when they are based on only a few items (Pommerich, Nicewander, and Hanson, 1999). Depending on how the results are to be reported, this issue might be exacerbated with the use of an augmented shelf-assessment (i.e., if scores are derived for created objectives or sets of small items across both the shelf and augmented portion of the assessment). Perusal of a few technical manuals (for both off the shelf and customized assessments) show reliabilities for part-scores too low to adequately infer student level performance, and these manuals often have clauses cautioning against over-interpretation of such scores. This is true for both shelf and custom created standards-referenced assessments. This should not be a surprise and is a simple artifact of the relatively few number of items or score points contributing to the part-score. However, despite what Cook (1950, pg. 29) and his colleagues refer to as "...other tests of a more diagnostic nature..." diagnostic or "actionable" results are still desired from these part-scores as teachers, parents and students are arguably likely to use such scores to focus remedial efforts. Also not surprising, upon retest a completely different profile might emerge potentially leaving the test-taker confused and likely to "blame the test" for being inconsistent.

In order to counter potential misinterpretation or over-interpretation of unreliable part-scores, some statistical procedures have been investigated. To be useful in guiding student learning, resulting scores for sub-domains are more valuable the more specific they become (Wainer et. al., 2000). The more specific they become, however, the less reliable they typically are. This is the motivation for statistically adjusting or improving the reliability of part-scores. Various methods in this regard have been derived and investigated (Yen, 1987; Yen, et. al., 1997; Bock, Thissen, and Zimowski, 1997; Pommerich, et. al., 1999; Wainer, et. al., 2000; Gessaroli, 2004; Kahraman & Kamata, 2004; Tate, 2004; Shin, Ansley, Tsai and Mao, 2005). Most of these procedures use collateral test information in order to enhance the reliability of the part-scores. If such scores can be made more reliable, even if only statistically, then perhaps

valid and useful links to instruction can be made. Given the availability to collect collateral information from an augmented shelf-test (norms, other subscale scores, potentially multiple content areas); such research would seem to be very promising for augmented assessments.

Another problem with the usefulness of part-score information is translating performance levels (i.e., proficiency levels under NCLB) down to the part-score level. The logic is straight forward; a student failed to achieve proficiency on the assessment and did poorly on a particular objective. How much do they need to improve on that objective to be proficient? Currently, many states (California, Arizona, Michigan, and Tennessee to name a few) use procedures to perform this estimation or projection of cut scores to the subscale level. One of the more widely used procedures includes IRT techniques relating total test ability (IRT theta values) to the part-score subscale. Another method uses a more a norm-referenced approach approximating average student performances on the overall test with that of the subscale. In the augmented shelf-test arena this raises several questions: should this approximation be to the shelf-portion only, the augmented or additional items only or to the total score from both sets of items?

Often, part-scores are used to create student learning profiles, which are really just patterns of scores across subscales. Such profile scores have been depicted graphically (Mehrens & Lehmann, 1973, pg. 160):

“...a graphic representation of results on several tests, for either an individual or a group, when the results have been expressed in some uniform or comparable terms (standard scores, percentile ranks, grade equivalents, etc.). The profile method of presentation permits identification of areas of strength and weakness.”

Hills (1993, pg. 26) also stressed the use of graphical representation, and described a profile as a collection of test scores or objective scores place on: “...a graph or chart, side by side, using the same scale for all of them so that comparisons can be visualized.”

This notion from Hills is not so different from what Cook and his colleagues almost 60 years ago desired or claimed were needed. Unfortunately, if the profile requires the part-scores to be on the same scale or use comparable metrics, their usefulness in the augmented shelf-test arena is limited. Such part-scores from an augmented NRT are not likely to be on the same scale—this is not impossible but the efforts, assumptions and requirements involved might render the resulting augmented assessment more expensive and complicated to build than a new test built from scratch. This is speculation of course, but speculation that can be confirmed

empirically. Frankly, the entire concept of placing both the shelf and augmented items onto the same scale poses technical challenges by itself regardless of its fit for use in profile analysis.

User Acceptance and “Usability”

Despite the arguments made in this paper that NCLB results are primarily intended for program evaluation and not useful as a tool to help plan individual student instruction, teachers nonetheless are looking for this kind of information. Huff and Goodman (2007) presented the information in Table 1 regarding how teachers who received statewide mandated results used them. Given that most large-scale NCLB-type assessments are given late in the spring and contain relatively few items (depending upon grade and subject this could be as few as 30-40 for reading and 40-60 for mathematics); Table 1 suggests a desire by teachers to use the assessment results, in some way, to inform instruction on a daily basis.

Similarly, Hagge and Waltman (2008) found that a large percentage of teachers (in elementary school this was about 80% of the teachers in the study), attributed increases in the use of data, attention paid to lower performing students, and opportunities for professional development to NCLB. Clearly, results matter for instruction at least as it applies to NCLB.

<b>How Results were Used</b>	<b>Percentage of Teachers Receiving Results</b>
Use results daily to inform instruction	31%
Use results a few times a week	14%
Use results a few times a month	15%
Use results a few times a year	21%
Use results only one time a year	14%
Never use results	7%

**Table 1. How Teachers Use Assessment Results.**

Pearson’s own work (Meyers, Shin & Nichols, 2008) on what features of an “instruct-assess-report” feedback system integrated within instruction was also not surprising. Usability experts facilitated face-to-face focus group meetings and conducted phone interviews with not only teachers, but principals, parents and test coordinators as well. This research, following the

principles of user-centered design, found the perception that three features of a system to support instruction were required: rapid return of results, reports tailor to the audience and the availability of or access to instructional interventions.

Trout and Hyde (2006) found similar results from their research conducted across multiple states. They found that teachers reported having little time to view actual student reports, suggesting that graphical displays and report efficiency was important. They also reported that while colorful reports were impressive, color alone was not the driving or primary focus. Teachers embraced the idea of dynamic reporting via the web and like the concept of tailoring reports based on the needs of the audience.

Given these rather consistent teacher perceptions about the value of reports (i.e., timeliness and customizability to the audience), questions regarding the reporting of a mixture of shelf and customized assessments can be asked. For example, if reports need to be customized for the audience or dynamic in order to be instructionally relevant, will the development of such reports for an augmented shelf-test eliminate the economies of scale implicit in the value of the augmented assessment? Additionally, if teachers have little time to review existing reports, will the possibility for three sets of such reports (shelf-test only, augmented only and aligned shelf items plus augmented items) be too much for teachers to use even if it is online and dynamic? Clearly, more research regarding the formatting and presentation of reports for augmented assessments is required.

### **Conclusions**

This paper looked at the issue of score reporting from shelf-tests under NCLB. It started as a premise that any shelf-test had to be augmented in order to fulfill the requirements of NCLB. Arguments were presented that such augmentation may modify the construct definition of the assessment and that differences in the sampling of the defined domains (shelf-test, augmented and combined) potentially threatens the validity of resulting score interpretations. This argument was supported based on circumstantial evidence, logical deduction as well as research evidence from the late 1980s and early 1990s.

A historical perspective revealed that few of the desires delineated over 50 years ago in measurement have been fulfilled when it comes to score reporting. It was argued that score reporting and the inferences generated from the scores should support individual learning and that such inferences are required by the technical Standards. The paper argued, however, that the

primary goal of NCLB reporting is program accountability and not individually delineated steps for instructional improvement. These arguments applied regardless of the type of assessment, (shelf, augmented or standards-referenced).

Two ways to enhance the validity of part-score reporting were briefly discussed. It was argued that one of the most promising techniques for improving the reliability of part-scores may be found in statistical procedures using collateral test information. Given the potential for the increased amount of collateral information from augmented shelf-tests, this aspect of current research seems promising. More research is needed regarding how these techniques apply to the augmented shelf-test arena.

Finally, the need for additional research regarding the fundamental principles of information processing, usefulness and usability was discussed. Current research shows that teachers desire actionable or instructionally relevant information from assessment, yet have little time to review or customize reported results themselves. Thus, research that addresses end-user input, instructional design and cognitive processing to understand better the way to convey score information in an efficient manner is needed and would be highly desired.

## References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Bock, R. D., Thissen, D. & Zimowski, M. F., (1997). IRT estimation of domain scores. *Journal of Educational Measurement*, 34, 197-211.
- Cook, Walter W., (1950). The functions of measurement in the facilitation of learning. In Lindquist, E. F., Editor, *Educational Measurement*. American Council on Education, Washington, D. C.
- Cruse, K. L., & Twing, J. S., (2000). The history of statewide achievement testing in Texas. *Applied Measurement in Education* 13(4), 327-331.
- Education Week (2006). NCLB: Act II—The latest news on the reauthorization of the No Child Left Behind act. Retrieved from the web, March, 2008:  
[http://blogs.edweek.org/edweek/NCLB-actII/2007/12/growth\\_models\\_for\\_all\\_who\\_qual\\_1.html](http://blogs.edweek.org/edweek/NCLB-actII/2007/12/growth_models_for_all_who_qual_1.html).
- Forsyth, R. A., Twing, J. S., & Ansley, T. N., (1992). Three applications of customized testing in local school districts. *Applied Measurement in Education*, 5(2), 111-122.
- Gressaroli, M. E., (2004). *Using hierarchical multidimensional item response theory to estimate augmented subscores*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.
- Hagge, S. L. & Waltman, K., (2007). *Teacher perceptions of NCLB: A multi-year study*. Distinguished Paper—Iowa Educational Research and Evaluation Association annual meeting, Iowa City, Iowa.
- Harris, D. J. (1987). *Estimating examinee achievement using a customized test*. Paper presented at the meeting of the American Educational Research Association, Washington, DC.
- Huff, K. & Goodman, D. P., (2007). The demand for cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 19-60). New York, NY: Cambridge

- Kahraman, H., & Kamata, A., (2004). Increasing the precision of subscale scores by using out-of-scale information. *Applied Psychological Measurement, 28*(6), 407-426.
- Lindquist, E. F., (1950), (Editor), *Educational Measurement*. American Council on Education, Washington, D. C.
- Linn, R. L., & Hambleton, R. A., (1992). Customized tests and customized norms. *Applied Measurement in Education, 4*(3), 185-207.
- Meyers, J. L., Shin, D., & Nichols P. D., (2008, January). *Perspective: An integrated assessment and instructional resources system*. Pearson Research Report. Iowa City, IA: Pearson.
- Pommerich, M., Nicewander, W. A., & Hanson, B., (1999). Estimating average domain scores. *Journal of Educational Measurement, 36*, 199-216.
- Shin, C. D., Ansley, T., Tsai, T., & Mao X. (2005). *A comparison of methods of estimating objective scores*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.
- Smisko, A., Twing, J. S., & Denny, P., (2000). The Texas model for curricular validity. *Applied Measurement in Education, 13*(4), 333-342.
- Tate, R. L., (2004). Implications of multidimensionality for total score and subscore performance. *Applied Measurement in Education, 17* (2), 89-112.
- Trout, D. L. & Hyde, B., (2006). *Developing Score Reports for Statewide Assessments that are Valued and Used: Feedback from K-12 Stakeholders*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.
- Twing, J. S., Forsyth, R. A., & Ansley, T. N. (1990, April). *An application of customized testing with local school populations*. Paper presented at the annual meeting of the NCME, Boston.
- Wainer, H., Vevea, J. L., Camacho, F., Reeve III, B. B., Rosa, K., Nelson, L., Swygert, K. A. & Thissen, D. (2000). Augmented scores—“borrowing strengths” to compute scores based on small numbers of items. In D. Thissen, & H. Wainer (Ed.), *Test scoring*. (pp. 343-387). Hillsdale, NJ: Erlbaum Associates.

Way, W. D. (1988). *An evaluation of item response theory ability estimates obtained for local school populations using customized achievement tests*. Unpublished doctoral dissertation, The University of Iowa, Iowa City, IA

Way, W. D., Forsyth, R. A., & Ansley, T. N. (1989). IRT ability estimates from customized achievement tests without representative content sampling. *Applied Measurement in Education*, 2(1), 15-35.

Yen, W. M., (1987). *A Bayesian / IRT index of objective performance*. Paper presented at the annual meeting of the Psychometric Society, Montreal, Quebec, Canada, June.

Yen, W. M., Green, E. R., & Burket, G. R. (1987). Valid normative information from customized achievement tests. *Educational Measurement: Issues and Practice*, 6, 7-13.

Yen, W. M., Sykes, R. C., Ito, K., & Julian, M. (1997). *A Bayesian / IRT index of objective performance for test with mixed-item types*. Paper presented at the annual meeting of the National Council on Measurement in Education in Chicago.