Improving the Information Value of Performance Items in Large Scale Assessments

P. David Pearson
Diane R. Garavaglia
Michigan State University

Commissioned by the NAEP Validity Studies (NVS) Panel
August 1997

George W. Bohrnstedt, Panel Chair
Frances B. Stancavage, Project Director

The NAEP Validity Studies Panel was formed by the American Institutes for Research under contract with the National Center for Education Statistics. Points of view or opinions expressed in this paper do not necessarily represent the official positions of the U. S. Department of Education or the American Institutes for Research.

The NAEP Validity Studies (NVS) Panel was formed in 1995 to provide a technical review of NAEP plans and products and to identify technical concerns and promising techniques worthy of further study and research. The members of the panel have been charged with writing focused studies and issue papers on the most salient of the identified issues.

For Information:

NAEP Validity Studies (NVS)
American Institutes for Research
1791 Arastradero Road
PO Box 1113
Palo Alto, CA 94302
Phone: 415/493-3550
Fax: 415/858-0958

Acknowledgments

We are most grateful to all of the members of the National Validity Studies Panel for their helpful comments throughout the development of this report. We are particularly grateful to our panel advisors—John Dossey, Illinois State University, Richard Duran, University of California, and Robert Linn, University of Colorado—for their patience and continuous support and feedback. From the American Institutes for Research, both George Bohrnstedt (panel chair) and Fran Stancavage (project director) provided much support and many suggestions for polishing the manuscript.

Finally, Michael Rodriquez, here in our offices at Michigan State University, read and responded to several versions along the way and pointed us toward some critical references.

Contents

The purpose of this essay is to explore both what we know and what we need to learn about the information value of performance items *when they are used in large scale assessments*. Within the context of the National Assessment of Educational Progress (NAEP), there is substantial motivation for answering these questions. Over the past decade, in order to adequately portray the breadth and depth of important curriculum standards, NAEP designers have invested substantial time and energy in creating extended constructed-response items and enormous financial resources in scoring these items. While these items are popular with curriculum experts within various content areas (Linn, Glaser, & Bohrnstedt, 1997), it is not clear whether they possess the marginal utility required to justify their cost; that is, they may not provide new information above and beyond that which is provided by a more standard mix of multiple-choice and short items with known measurement characteristics and much more economical scoring protocols. Even worse, there is some reason to believe, based on research in non-NAEP settings, that extended constructed-response items may provide negative returns in terms of the overall goal of accurate measurement of performance on some broad domain such as reading, mathematics, or science (see Forsyth, Hambleton, Linn, Mislevy, and Yen, 1996).

To raise this issue is not to question the inherent validity of these tasks within internal assessment systems (i.e., as they are embedded within particular curricula), but only whether they have a role in the "drop-in-from-the-sky" assessment approach of external assessments such as NAEP (Forsyth, et al, 1996). In order to prove worthy of the added design and scoring costs in the latter context, however, they must be shown to play a significant role in improving the accuracy and/or the credibility of these externally motivated large-scale assessments.

A more specific version of this broad question is contained in the initial charge provided to us by the panel overseeing a set of validity studies to guide the future development of NAEP (AIR, 1995):

> This paper would review what is currently known about the psychometric properties of performance items in large scale assessment generally, and in NAEP in particular. The paper would then go on to propose modifications that appear to hold the greatest potential for improving the information value of such items and to lay out the design, time line, and budget for one or more empirical studies to explore (a subset of) these modifications. Possible candidates for study might include ways of obtaining multiple data points from single exercises, refining item directions and scoring rubrics to improve alignment with content framework constructs, improving the different characteristics of such items, etc. The costs of large scale administration and scoring will be taken into consideration in evaluating alternatives for further study.

To dispatch our charge, we have traversed the "information value" terrain along as many paths as we could find—combing the measurement literature to determine the various ways in which scholars have conceptualized and operationalized the "information value" construct, reviewing research conducted within those approaches, consulting essays that emphasize the importance of strong conceptual grounding in content frameworks,

studying the broadest available construal of the construct of validity (e.g., Messick, 1989), and, finally, and most unsatisifyingly, attempting to determine, at a conceptual and philosophical level, what the assessment community means when they talk about the information provided by assessments.

We have organized our essay into four sections. First, we consider the construct of information value in its broadest philosophical sense and then describe the classical ways of operationalizing it. Second, we review the available literature within each of these operational traditions (IRT, factor analysis, correlational studies). Third, we consider some alternative versions of information value, based more on cognitive, conceptual, and pragmatic considerations. Finally, we outline a series of studies that we believe ought to be supported by the National Center for Educational Statistics (NCES) in order to answer the question of how NAEP can be modified to increase its "information value."

The Idea of Information Value

Information has "value" to the degree that it helps us make decisions or draw conclusions about issues or questions that matter to us. If we have only a single datum available to make a decision, then the question of value added (what marginal information do I get?) is moot because "what you see is what you get." But as soon as an additional datum is available, the question of value added becomes important, and the common sense question is whether we will make a better (i.e., more valid or more accurate) decision with this added information available to us.

When it comes to educational measurement, and in particular large scale educational measurement, the issues and questions that "matter to us" usually take the form of queries about cognitive processes or curriculum, such as,

- How well does this population read?
- What does this population know about mathematical problem solving?
- How much does this population know about American history?

A datum, or bit of information, within a psychometric framework could be a score on a test or a score on a test item. The question of interest, vis-à-vis information value, is whether an additional bit of information helps us better answer the question(s) of interest. Ultimately, however, the issue of how new information adds value must be addressed, and the answer is not always straightforward. New information can serve several functions in the decision-making processes of individuals or groups.

1. First, new information could increase our *confidence* in the decision we have already made or are about to make. Psychometrically, this is tantamount to increasing the psychometric *precision* of our measurement. Within the par-

lance of Item Response Theory (IRT), this is typically what is indexed by the information function. By ascertaining the item information function for groups of items, we can make decisions about the ability level on the underlying trait at which particular sets of items provide maximal information (or, if you prefer, maximal capacity to discriminate among individuals with different levels of the ability in question). IRT analysis has occasionally been employed to compare the relative information contributions of multiple-choice and constructed-response items. For example, Bridgeman (1992) examined the item characteristic curves (ICCs) for mathematics items (on the GRE quantitative scale) that appeared in both multiple-choice and constructed-response (filling in spaces on a grid) format. He found that the multiple-choice items were generally easier but the relative difficulty of paired items (one mc item and one cr item measuring precisely the same content) varied unpredictably across ability levels.

2. "New" information focuses on *novelty* rather than precision. In other words, the additional information leads to the elaboration of a new construct or dimension. Thus, instead of increasing the precision of our measurement of a particular construct, such as mathematical power, an additional item, subtest, or format could give us information about a related construct, such as mathematical problem solving. Several studies have been conducted to determine the value added in this sense of new information. Factor analysis is the most common tool used to calibrate the contribution of hypothetically separate items or item sets (do items in different formats load on different factors?), although occasionally researchers use correlational or regression analysis to examine the relative distributions of common and unique variance among different sets of items. For example, Ackerman and Smith (1988) examined the relationships of several multiple-choice and direct forms of writing assessment and concluded that a six-factor model, corresponding roughly to the different forms of writing assessment included, provided the best fit to the data. By contrast, Ward, Dupree, and Carlson (1987), applying similar statistical analyses to reading comprehension tests, found only weak evidence for unique components, based either on format or cognitive demands.

3. New information could be thought of as more psychological than statistical in character. For example, additional items or item types could provide us with absolutely no new statistical information (i.e., neither an increase in the precision of test information about a given dimension nor any information about a new dimension) but still give us greater confidence that we had measured the domain of interest adequately. This added confidence could stem from one of two sources, both of which are conceptual rather than technical. In the first instance, trust is at issue. Suppose we have a complex domain, such as reading comprehension, which is known to have several aspects or dimensions, such as literal, inferential and critical stances toward a text. A small (representative) sample of items from the domain might leave us with

the concern that not all of the aspects of the domain had been adequately represented, even though it provided as reliable an estimate of domain ability as a much larger sample would show. Increasing the sample of items would contribute to a greater sense of "trust" without any increase in technical information. In the second instance, the cognitive demands of the tasks are at issue. Snow (1993) makes this point vividly in commenting about the relationship or lack of relationship between psychometric and psychological equivalence:

> . . . a constructed-response test and a multiple-choice test may correlate in some student population about as high as their respective reliabilities will allow; this fact may permit the two tests to be considered psychometrically equivalent for use in rank ordering students . . . But the two are not psychologically equivalent; we only act as if they were . . . (p. 46).

Implicit in Snow's assertion is an assumption that in order for two tests (or items) to be psychologically equivalent, they must engage test-takers in the same cognitive and/or affective processes. The discontinuity between psychometric and psychological constructs is not a trivial matter. Ultimately it is a question of construct validity, and, as such, involves an examination of the validity of the original framework used to characterize the construct, the practices used to translate the framework into test items, and the uses of information obtained from the test.

4. New information could arise from taking a second perspective on data or tasks that had already been examined from a different perspective. For example, suppose an essay written for a social studies, science, or mathematics performance was examined initially with a rubric designed to calibrate the amount of specific, subject matter-learning demonstrated by students. On a second pass, the same essay could be examined with a writing rubric measuring accomplishment on writing dimensions such as voice, power, audience, and mastery of conventions. The additional data generated by the second scoring could be "useful" and provide evidence for making an additional decision even if it were *not* statistically independent of data from the first scoring. For example, in the work of New Standards (Myers & Pearson, 1996), students participated in three- or four-day integrated language arts tasks in which they read texts, responded to them, discussed issues in groups, and, finally, wrote in response to specified prompts. The culminating writing responses were scored first for writing power and then for evidence of students' ability to synthesize information from text. Similarly, the Maryland performance assessment (1992) allows for a single set of responses to be scored initially for science, mathematics, or social studies and later for either writing prowess or reading comprehension. A variant of the second scoring approach is the tradition of dimensional or primary trait scoring currently-

popular in writing assessments. The idea is to make several passes through each constructed-response item or paper, each time scoring it holistically (i.e., considering the whole essay) but for a different dimension.

5. Information could be improved in a pragmatic sense without any increase in the quality of the psychometric information or even an increase in the quality of information about the underlying psychological construct. New information could provide us with a new perspective on a particular decision. For example, in analyzing the results of a mixed (performance and multiple-choice) battery of geography assessments, Kon and Martin-Kniep (1992) wanted to know, "Do we learn things about students' knowledge of geography and ability to use geography skills that we would not learn using other testing methods?" Their criterion for learning new things derived from common sense (is this a new type of task or activity?) and weak statistical comparisons (are the two types of items reasonably unrelated, as indexed by correlation coefficients?). In a study of mathematics performance cited earlier, Bridgeman (1992) concluded that while providing no new information about overall performance in the domain (there was a high degree of common variance), the gridded items could provide readily information about specific skills within the domain or about dominant error patterns (should anyone care to create subscale scores or conduct error analyses). The arena for determining whether a test provides new or better information by this criterion is neither psychometric nor psychological analysis; it can only be assessed within a context, usually instructional, in which people use a variety of assessment information to make real decisions of consequence for others. It might turn out, for example, that constructed-response items, or the factor on which they load, are correlated at unity with multiple-choice items, but so dramatically increase the confidence and/or quality of decisions that they constitute an indispensable piece of a broader "assessment system." In everyday parlance, this might be equivalent to feeling confident that the individuals who were being assessed both *possessed* and *could apply* certain domains of knowledge.

The Information Value of Performance Assessments

Within the corpus of wide-scale research on the constructed-response/ multiple-choice relationship, the question of interest has been whether constructed-response items, when used with multiple-choice items in a mixed format package, add "value" (i.e., additional information) to processes of understanding, reporting, and using assessment results. For our purposes, it is more informative to decompose the overall question into a two part question: a) do constructed-response items provide us with more information about what students are capable of doing than we would get from multiple-choice items alone, and, if

so, b) what types of skills are tapped by the constructed-response items that are not measured by multiple-choice items?

Much of the archival literature evaluating the value added of performance items arises from analyses of data from the College Board's Advanced Placement (AP) exams. This should not be surprising since the AP exams have a long history of using both sorts of items, presumably because designers and users believe that not all higher order skills can be assessed easily with multiple-choice items.

Working within an IRT framework, Lukhele, Thissen, and Wainer (1994) found that constructed-response items on the test provided little information beyond that which multiple-choice items yielded. This result was the same for both the chemistry and U.S. history exams analyzed in that study. In fact, for the chemistry test, twice as much information was yielded from 16 multiple-choice items compared to one constructed-response item when a three-parameter logistic IRT model was utilized. Of course, given what we know about the influence of test length (as indexed by the number of items) on information value (e.g., Hambleton & Swaminathan, 1985), it should not be surprising that a 16-item test provides more information than a one-item test. Nevertheless, when test length is indexed by total testing time, the 16 multiple-choice items are equivalent in length to the one constructed-response item and cost much less to score, indicating that from a cost-effectiveness perspective, the multiple-choice items are preferable.

Additional studies involving AP exams in chemistry, science, and computer science (Thissen, Wainer, & Wang, 1994; Wainer & Thissen, 1993; Wainer, Wang, & Thissen, 1991; Wang, Wainer, & Thissen, 1993) have further explored the value added question. The evidence from all of these studies suggests that when data from performance items are combined with data from multiple-choice items, little new information about the skills being assessed is added.

In other content areas, however, some evidence to the contrary has been uncovered. In the domain of writing, Werts et al. (1980), working with first-year college students, attempted to determine whether different item formats on six tests would uncover different writing traits. The design was a variation of multitrait-multimethod. The instruments included three administrations of the Test of Standard Written English (TSWE) and three, short (20 minute) essay prompts. All of these tests were given within the same year. The nonzero covariation among the essay residuals showed that the essays measured some common trait that was different from whatever traits the essays and TSWE shared. One positive aspect to this study was the independence of the three essay items in that they were obtained from three different writing occasions. In the Ackerman and Smith (1988) study reported earlier in this report, evidence of unique format contributions also were found.

Bennett et al (1990) used confirmatory factor analysis to examine the infrastructure of the AP computer science examination. Each constructed-response item was treated as a separate variable; groups of ten or more multiple-choice items were formed, with each

group representing a separate variable. A one factor covariance structure model was used to analyze the data. The results indicated that both item formats measured the same characteristics, suggesting that the addition of the constructed-response items did not provide any additional information about the trait under investigation. In a later study by Bennett et al (1991), however, which used a structure hypothesizing separate format factors, the investigators found that the disattenuated correlation coefficients were significantly (though not substantially) different from unity.

Even though added information for mixed item formats has been demonstrated occasionally (Ackerman & Smith, 1988; Werts, et al, 1980), it has not been possible to specify the different characteristics or skills measured by different item formats. Further, the evidence is generally shaky because of methodological weaknesses in the research. For example, some constructed-response items were generated through simple transformations of existing multiple-choice items rather than through any careful analysis of the types of tasks that might best be measured in the constructed response format. Also, the small number of constructed response items used in this work plays havoc with the interpretation of conventional statistical analyses (Mazzeo, Yamamote, & Kulick, 1993). Finally, even when evidence indicates that different item formats appear to assess psychometrically distinct traits (e.g., in the Ackerman and Smith study), without knowing precisely what the item types actually assess we cannot say whether they truly measure different traits or simply measure the same underlying traits differently. This concern points to the importance of having a clear and well-articulated model of the construct in question as a guide to item development and psychometric analysis.

Issues and problems with factor analytic studies

Factor analysis has been used in several of the studies that have attempted to differentiate and uncover exactly what is being assessed by different formats. While robust with respect to many assumptions and issues, factor analysis is not without problems. First, results and interpretation of the results may depend on the model selected for the analysis, as well as the number of factors included in the model. Traub (1993) found that when a one-factor model was used to fit the correlation coefficient matrices, the results showed that the two-item formats actually measure the same characteristic. When the same data set was analyzed using a two-factor model—one factor for each format—the results indicated that the coefficients for the two factors were significantly, albeit only slightly, less than one (Traub, in Bennett and Ward, 1993).

Second, the evidence is equivocal, sometimes even within the same study. Using a hierarchical factor model that included a general factor for all items and two orthogonal factors for the constructed-response items, Thissen (1994) found that constructed-response items produced a statistically significant factor, orthogonal to the general factor, suggesting that they measure something uniquely different from the multiple-choice items, which loaded almost exclusively on the general factor. The constructed-response items also loaded on the general factor; more importantly, their loadings on the general factor were usually larger than on the constructed response factors. This pattern of results

suggests the constructed-response items are measuring the "same" content as the multiple-choice items but are measuring something additional, which may be format. These results might be construed as evidence that what is being assessed by either item format is probably being assessed poorly; furthermore, we do not know for certain what constitutes the "thing" that *is* being measured differently be each item format.

Also problematic is the consistent difference in the difficulty level of multiple-choice and constructed-response items. Thus, what appears to be construct unequivalence may be nothing more than difficulty loadings. These concerns point to the importance of basing psychometric analyses on well-articulated theories of the constructs being measured. Otherwise, when the factor analysis suggests separate factors, it will provide little or no guidance about what those factors represent.

Third, study design can cloud results and interpretation when using factor analysis. The most common design flaw in these factor analytic studies is the small number of constructed-response items uses for analysis. While small numbers of items do not inherently lead to unreliability, the capacity to find additional factors when they really exist is compromised by small item samples. The question is whether we are giving the constructed response format a chance to demonstrate uniqueness if and when it really exists. Also the way the constructed-response question is scored may have an impact on results; others things being equal, dichotomous holistic scores are likely to provide less information than a set of dimensional scores for the same item.

Goldstein (1994) points out a fourth problem with traditional exploratory factor analytic techniques; this issue has to do with the "reference population." When more than one subpopulation is under study, it is possible for the model to "fit" well in one subgroup but not in the other subgroup(s), e.g., ethnic or gender difference studies. The ideal solution is to conduct separate analyses for each subpopulation using confirmatory factor analysis; however, subgroup analyses can lead to a host of reliability problems if there are small samples in each subgroup. This is especially problematic in analyses of race and ethnicity; usually the sample size for Caucasian and African American students is adequate, but analyses of other ethnic groups are risky at best.

Refocusing our energies

Some scholars of assessment (for example, Cronbach, 1988; Nickerson, 1990; Snow, 1990; 1993) have encouraged us to set aside our psychometric and methodological approaches to studying validity issues and turn our attention instead to the constructs underlying the assessment enterprise and to a "psychology of test design" (Snow, 1993). Snow, in particular, has encouraged us to expand our "typical" way of thinking about construct validity so that we can escape the boundaries of our current paradigms.

The goal of research stemming from this perspective is to uncover the psychological underpinnings of a particular domain and then to relate them to the psychological

behaviors involved in answering certain items. Snow also encourages us to expand our thinking and research beyond the obvious cognitive demands of assessments:

> Although cognitive analysis of the contrast between constructed-response and multiple-choice test formats is essential, so is analysis of the conative (i.e., motivational-volitional) and affective aspects of performance that connect to the contrast (p. 46).

The key is to recognize not only that different item formats may elicit different cognitive characteristics (structures, to borrow from information processing) or processes which are necessary to respond successfully to the different item formats, but that the interpretation of scores also depends on whether different item formats elicit unique motivation, effort, or consequential behaviors or dispositions. For example, consider the following questions:

- Do students exert the same amount of mental effort when they answer a multiple-choice item as when they answer a constructed-response item?
- In comparison with the panic that arises when students encounter a constructed response item for which they have absolutely no idea of even where to begin, are students more likely to take a risk and guess at the answer to a multiple-choice item in the belief that they have some non-zero chance of succeeding?
- Does confidence in selecting (or even guessing at) an answer differ markedly from confidence in one's ability to express oneself in writing?

Snow (1993) goes on to propose an approach for investigating whether affective and conative characteristics are at play, along with cognitive processes. Building on the work of Cronbach (1988), he suggests that we adopt a "rival hypothesis" approach to studying these issues. To quote Cronbach:

> The advice [to pursue rival hypotheses] is not merely to be on the lookout for cases your hypothesis does not fit. The advice is to find, either in the relevant community of concerned persons or in your own devilish imagination, an alternative explanation of the accumulated findings; then to devise a study where the alternatives lead to disparate predictions. (p. 14)

Then, of course, one would collect the data that would allow evaluation of the rivals. Snow suggests an additional protective guideline—that the investigator word the hypothesis in a way that contradicts his or her personal belief(s). For example, an advocate of constructed-response format would phrase the hypothesis to favor a multiple-choice format, and vice-versa.

New Initiatives

Even though there exists a small corpus of careful studies that allow us to examine the relationship between multiple-choice and constructed-response items, we still have a

great deal to learn. Much of the problem in interpreting the current set of studies is that the research has been more opportunistic than intentional. In the prototypic study, researchers take advantage of the fact that an existing test or battery happens to have included both constructed-response and multiple-choice formats. Much rarer are studies in which the researchers have set out to design, from scratch, studies that have, as their expressed purpose, the evaluation of both the underlying constructs and the validity of the test(s) designed to measure those constructs. Messick (1993, p. 64) recognized this problem in commenting upon our tendency, whether we come from a psychometric or a psychological perspective, to work from what we have rather than from a concept of what we want to learn: "However, both perspectives tend to rely too heavily on the construct analysis of existing tests, whether by factor analysis or task analysis, rather than focusing on theories of the construct domain as a guide to designing construct relevant tests."

What is needed is a fresh examination of the relationships between multiple-choice and constructed-response items—an examination that begins with an explication of a theory of the domain being assessed, which is then transformed into a theory of achievement within the domain. Such a theory would ultimately have to extend beyond the domain into more generic measurement issues (e.g., item format, test length, assessment context) and motivational issues (e.g., stakes, examinee preparation and anxiety) in order to test predictions about the nature and strength of relationships among components and/or between each component and some external criterion measure of the domain. This is, of course, exactly the sort of activity that psychometricians have had in mind for years in discussing the construct validation of tests (Messick, 1989; Cronbach, 1971). It is also, an activity that, for a variety of reasons, has escaped our attention or exceeded our capacity; we seldom approach that ideal. Nevertheless, short of a complete evaluation of the construct in question, there are a number of useful but less ambitious initiatives that would allow us to answer the question of value added for performance items with greater assurance than is currently possible. As first steps at working toward building this theory, we close this paper by sketching out, in broad terms, a set of studies that should provide us with needed information about the value added of performance items in mixed format assessments such as NAEP.

The cognitive demands of multiple-choice and constructed-response items

Even if it were to be demonstrated that comparable sets of multiple-choice and constructed-response items were psychometrically equivalent, it does not follow that they are psychologically equivalent, i.e., that they are variant indices of the same underlying construct; high degrees of shared variance could stem from any number of circumstances. We need studies in which we examine psychological equivalence directly. Even better would be studies in which we examine psychometric and psychological equivalence for a common set of items. The most productive tool for determining the psychological processes elicited by test items is the think-aloud procedure, which has been used with NAEP items in some previous work (Yepes-Bayara, 1996; Campbell, 1996). Participants could be asked to share their step-by-step thinking as they attempt to select or construct responses to test items. Although think-aloud methodology appears promising for this

sort of initiative, it is by no means the only index of cognitive functioning that we should consider. When tasks involve text reading and response, both eye-movement methodology and computer controlled text search (look-back) methodology could tell us a great deal about the influence of item format on the role of text in selecting/ constructing responses.

The value of different types of items in educational decision-making

The question of whether different formats provide "additional information" is as much a question of "consumer" (i.e., test user) perception as it is psychometric independence. We need to know whether the addition of information from performance assessment provides new information *from the perspective of those who use the information for making decisions*, either about individuals or groups (in the case of NAEP, only decisions about groups are relevant). Ultimately, the question of whether constructed-response assessments possess interpretive value is an empirical question, and deserves to be answered empirically by observing the uses to which test users put the information they receive. We propose a study in which subject matter specialists are provided with tailored reports of NAEP results and asked to draw conclusions about student performance in their subject area (or, in the case of state-by-state comparisons, about the relative distribution of achievement). Within such a context, systematic variation in the nature of the information provided could be introduced (only multiple-choice, only constructed-response, or both). We would examine the influence of different information packages on the nature of the conclusions that these specialist draw about aggregate performance and the suggestions they make about changes in curricular or instructional policy.

Rubric research

The NAEP rubrics for reading have been roundly criticized by two separate evaluation panels. They are viewed as too quantitative and only marginally related to the NAEP framework for reading (DeStefano, Pearson, & Afflerbach, 1997). High dividends might result from a modest investment in creating new rubrics that are driven by the framework and then comparing the quality of information, both psychometrically and pragmatically, received when items are scored by these rubrics in contrast to the conventional rubrics. In another vein, we might examine the conceptual genesis of rubrics, paralleling Fredericksen's (1984) questions about whether transforming multiple-choice items into performance items is the same as transforming performance items into multiple-choice. Suppose the rubrics for a set of constructed-response items are based upon the same conception of underlying dimensions (the psychological construct) as were used to guide the development of a comparable set of multiple-choice items. Such a practice might, in fact, be reasonable if our goal is to examine trait equivalence across item formats; however, this practice can also constrain our thinking about the range of possible traits that might be assessed with the constructed-response format and teased out by an appropriate rubric. In other words, in achieving control for conceptual equivalence, we might be losing our capacity to uncover a larger set of possible dimensions of the construct that can only be tapped by the constructed-response format. This issue could be

addressed in a study in which competing rubrics were developed and used to score a common set of constructed-response items. The first rubric would be developed using a framework that had been used to generate multiple-choice items and then extended to constructed-response items. The second rubric would result from a fresh perspective: subject matter experts would be asked to generate a framework and related rubrics for an assessment system consisting only of constructed-response items. The question of interest is whether the two rubrics would yield equivalent scores and or trait information.

Truth in advertising

Many educators, and ordinary citizens find it odd that students are unaware of the criteria by which their extended constructed responses will be evaluated. It would be useful to know, even in a drop-out-of-the-sky assessment such as NAEP, whether students perform better when they know what is expected of them. We would easily embed an experimental form into NAEP; in such a form, a student version of the NAEP rubric would be attached to every extended constructed-response item, perhaps with an initial introduction to the importance of reviewing these rubrics before answering the questions.

Double scoring

Since reading (as a medium of information delivery) and writing (as a mode of response) are often employed in assessments of other subject matters (especially social studies and science, and, to a lesser extent, mathematics), the possibility arises that we might achieve greater cost-effectiveness by selecting some constructed-response items for double scoring. As suggested earlier, there is precedent for this practice in the work of the state of Maryland and the New Standards project, and there has been interest in this possibility within NCES (White, 1997). One possibility is to take advantage of already existing items which seem to lend themselves to double scoring. The cost advantages here are clear; in fact, the documents required to conduct the study (student test forms) may already exist within NAEP archives. A second possibility starts with the assumption that double-scoring is likely to be more effective if the items are initially constructed with that purpose in mind. In other words, if item writers knew that an item would be scored for both history and reading, both mathematics and writing, or both reading and writing, they might construct it differently from the way they would for single-subject scoring. A remote, but intriguing possibility would be to compare the efficacy of double-scoring using items that have been serendipitously generated versus those generated intentionally. In either situation, it would be important to examine carefully the cost-effectiveness of double-scoring. If, as some other studies suggest, scoring costs dwarf item-development and administration costs, then double-scoring may offer little or no overall cost savings.

Dimensional/opportunistic scoring

There is a core body of research, mainly in writing assessment, which examines the efficacy of scoring items or papers separately for two or more dimensions within a domain.

That research needs to be extended into other subject matter domains to determine whether constructed-responses are capable of yielding separate, perhaps even independent, estimates of performance on different aspects of a subject matter. In mathematics, for example, it is possible that a given response could be scored once for computational accuracy, a second time for evidence of problem-solving, and a third time for communication prowess. In reading, could a single response yield independent evidence of more than one of the basic stances in the NAEP framework: initial understanding, developing interpretation, personal response, or critical stance? Heretofore, we have tended to classify items independent of the responses they yield, or, in the case of reading, we have classified each level of the rubric for a given item into a given stance category (a level 1 response is initial understanding while a level 3 or 4 is critical stance). We could develop a more opportunistic scoring system for constructed-response items, one that was capable of taking advantage of and giving credit for evidence of performance on a given dimension wherever and whenever it emerged, even if it emerged in situations where the item developers were least expecting it.

The biasing effect of surface features of language

Constructed response items put a premium on the expressive, especially the writing, abilities of students. In a writing assessment, scoring procedures that reward clear expression and that privilege dominant (i.e., standardized or conventional) forms of English is both understandable and even desirable. But what role, if any, should clarity and conventionality of expression play when writing is used as the vehicle, as the response format, to get at other knowledge and skill outcomes? More importantly, do constructed-response items have a depressing effect on the scores of students whose primary language is not standard English? For example, given responses that are identical in content but which differ systematically in their adherence to standard English, will the responses receive the same scores? If bias exists in unspecified scoring procedures, can it be reduced or eliminated through the use of clear scoring guidelines (remember, it is the ideas, not the quality or clarity of the expression, that count) and benchmark papers (high scoring exemplars which do not adhere to standard English). On the face of it, such a study may not seem to fit under the "quality of information" rubric, yet closer analysis suggests that quality of information *about particular populations* is exactly what is at stake here. We could easily embed such a study in a regular NAEP scoring conference. We could devise experimental responses in which we systematically altered adherence to features of standard English while holding content constant, and we could then determine how the competing versions fared during a normal NAEP scoring session. In the event that NAEP contractors balked at such an experiment (most likely on the ethical principle that it is deceptive vis-à-vis scorers), the study would have to be contracted out. In either case, it seems critical to know whether this sort of bias is present within our scoring systems. This question seems relevant to constructed-response items in all areas save writing.

The role of passage difficulty in reading assessment

The issue of passage difficulty in reading, particularly its potentially depressing effect on performance of students at the lower end of the performance continuum, has been emphasized in a number of recent reports (e.g., Forgione, 1996; DeStefano et al, 1997; NAE, 1997), and concerned scholars and policy makers have called for the production of easier blocks of NAEP reading items so that low-achieving students can at least "make it onto the scale," or in the language of information value, so that we possess more information about the performance of low-achieving students. If these more "accessible" blocks are created, and if we are thoughtful about how we design and generate items across blocks, we have an opportunity to determine whether response format (multiple-choice versus constructed-response) or passage difficulty (or some unique combination of the two) is responsible for the current low information yields of constructed-response items. It might be, for example, that students have a lot more to say when it is relatively easy for them to read, digest, think about, and even critique the texts they encounter. It might also turn out that difficulty interacts with achievement level in such a way that easy passages provide opportunities for low-achieving students to shine whereas hard passages provide just the challenge that high-achieving students need to get involved in the assessment.

The prospect of differential outcomes for different populations

We have not emphasized the importance of imposing an individual differences filter on any or all of the studies outlined thus far. We are aware of the importance of determining how each of these questions would be answered for different groups of students; however, we are also mindful of cost factors associated with increases in sample size and the reliability threats that arise when ample subsamples are not used. However, we are open to the possibility that one or another of these initiatives cries out for partitioning samples on some well-reasoned basis.

Prior experience

As with different populations, prior experience could serve as a filter or an independent variable in several of the studies outlined earlier. Prior experience has two possible realizations, one at the classroom/school level and one at the individual level. At the classroom/school level, it is instantiated as instructional experience (opportunity to learn). If we can be sure of the type of curricular emphases different students have experienced (e.g., emphasis on critical stance or response to literature in reading or problem-solving and communication in math), and if we can locate populations with different curricular histories, we can test constructed-response item performance under both optimal and suboptimal conditions: Do students who have learned what the items are designed to measure perform at high levels compared to students who have received other curricular emphases? It would be interesting, for example, to conduct the think-aloud study described earlier in sites which exhibit just such a curricular contrast. At the individual level, of course, prior experience is instantiated as prior knowledge, the

impact of which is well-documented in reading and writing assessment. As with the passage difficulty issue discussed earlier, it would be useful to know whether students provide more elaborate and more sophisticated responses to constructed-response prompts when they are quite knowledgeable about the topic at hand. To answer this question, we would have to obtain an independent measure of topical knowledge for the subject passage (in the case of reading), prompt (in the case of writing), or task (in the case of science or math).

Transforming items across formats

When evaluating the equivalence of constructed-response and multiple-choice items, researchers sometimes begin with one set of items, say multiple-choice, and rewrite them as constructed-response, or vice-versa. In this way, they attempt to control the content and focus of the items across formats. In other studies, there is no attempt to control for content and focus; instead researchers take advantage of the fact that an existing test happens to contain *some* multiple-choice and *some* constructed-response items. What we need are studies in which both multiple-choice and constructed-response items are developed in ways that allow each to "put their best foot forward." To our knowledge, Fredericksen (1984) is one of the few researchers to consider the possibility that we may be introducing a source of bias when, for example, constructed-response items are generated by transforming an existing set of multiple-choice items. He also is one of the few researchers to develop multiple-choice items from an existing set of constructed-response items. This study would extend the logic of his dual source approach to item generation. This could be accomplished with a procedure along these lines:

- Identify a domain of interest, such as reading comprehension, response to literature, mathematical power, etc.
- Identify one group of item writers with reputations for developing first-rate multiple-choice items; identify a second group with equally strong reputations for constructed-response items.
- Set each group to work on developing a set of items for the domain of interest.
- When each group is finished, ask them to exchange item groups and, as best they can, transform each multiple-choice item into a constructed-response item and vice-versa.
- Create matched item sets, balanced for content and format.
- Administer to students, and evaluate relationships between constructed-response and multiple-choice item subsets.

While a study of this magnitude, both in terms of item development and requisite sample size, would be expensive if conducted as an independent, stand-alone initiative, it could easily be integrated into the item tryouts conducted as a matter of course in the NAEP item development process. Furthermore, it seems to lend itself to several content domains, and the question is timely, where timeliness is indexed by conflict, tension, and

interest within the field, in math, science, social studies, and reading. Moreover, with appropriate design manipulations, it may even be possible to address the tension that continues to haunt subject specialists: Why do we continue to use psychometric tools, such as IRT, which assume unidimensionality when we believe that the domain is inherently multidimensional?

Concluding Remarks

We believe that the question of what constructed-response items add to the interpretation of NAEP results, although important in any era, is critical at this point in NAEP's history. The critics are lining up to recommend that constructed-response items be eliminated in the name of economy and precision. Until, and unless, it can be demonstrated that there are grounds, whether psychometric, conceptual, or pragmatic, for maintaining or expanding the emphasis on constructed-response items, their use will be questioned, and questionable. Consequently, we believe that NCES and the National Assessment Governing Board (NAGB), as the custodians of NAEP's efficacy and credibility, should do everything possible to see that studies such as those proposed in this concept paper are carried out with all deliberate speed. If programs of research such as these are not undertaken, then the decisions about the relative emphases on different item formats will be based upon considered opinion or political expediencies. Our nation's assessment centerpiece deserves a better hearing, one in which evidence is paramount.

References

Ackerman, T. A., & Smith, P. L. (1988). A comparison of the information provided by essay, multiple-choice, and free response writing tests. *Applied Psychological Measurement,* 12 (2), 117-128.

Bennett, R., Rock, D., Braun, H. Frye, D., Spohrer, J., and Soloway, E. (1990). The relationship of constrained free-response to multiple-choice and open-ended items. *Applied Psychological Measurement,* 14(2), 151-162.

Bennett, R., Rock, D., Wang, M. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement*, 28(1), 77-92.

Bennett, R. and Ward, W. (Eds) (1993). *Constructing versus Choice in Cognitive Measurement: Issues in Constructed Response, Performance Testing, and Portfolio Assessment*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement,* 29 (3), 253-271.

Cronbach, L. J. (1971). Test validation. In R.L. Thorndike (Ed.) *Educational Measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H.I Braun (Eds.), *Test Validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum Associates.

De Champlain, A. (1996). The effect of mulitdimensionality in IRT true-score equating for subgroups of examinees. *Journal of Educational Measurement,* 33(2), 181-201.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist*, 39, 193-202.

Forsyth, R., Hambleton, R., Linn, R., Mislevy, R., & Yen, W. (1996). *Design Feasibility Team: Report to the National Assessment Governing Board*. Washington, DC: National Assessment Governing Board.

Goldstein, H. (1994). Recontextualizing mental measurement, *Educational Measurement: Issues and Practice*, 12(1), 16-19, 43.

Hambleton, R. And Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Boston, MA: Kluwer Nijhoff.

Harris, D. (1993). *Practical Issues in Equating*. Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA.

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed-response, and examinee-selected items on two achievement test. *Journal of Educational Measurement*, 31, 234-250.

Maryland School Performance Assessment Program. (1992). *Sample Task and Scoring Tools: Grade 3: Science, Reading, Writing, Language Usage*. Baltimore, MD: Division of Instruction.

Mazzeo, J., Yamamoto, K., & Kulick, E. (April, 1993). *Extended Constructed Response Items in the 1992 NAEP: Psychometrically Speaking, Were They Worth the Price?* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta.

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

Messick, S. (1993). Trait equivalence as construct validity of score interpretation across multiple methods of measurement. In Bennett, R. & Ward, W (eds.), *Construction versus Choice in Cognitive Measurement* (pp. 61-73). Hillsdale, NJ: Lawrence Erlbaum Associates.

Myers, M., & Pearson, P. D. (1996). Literacy assessment in the New Standards project. *Assessing Writing*, 3, (1), 5-29.

Nickerson, R. (1989). New directions in educational assessment, *Educational Researcher*, 18, 3-7.

Snow, R. (1993). Construct validity and constructed-response tests. In Bennett, R. & Ward, W. (eds.), *Construction versus Choice in Cognitive Measurement* (pp. 45-60). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D., Wainer, H., & Wang, X-B. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice test? An analysis of two tests. *Journal of Educational Measurement*, 31, 113-123.

Traub, R. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In Bennett, R. & Ward, W. (eds.), *Construction versus Choice in Cognitive Measurement* (pp. 29-44). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ward, W. C., Dupree, D., & Carlson, S. B. (1987). *A Comparison of Free-Response and Multiple-Choice Questions in the Assessment of Reading Comprehension* (RR 87-20). Princeton, NJ: Educational Testing Service.

Werts, C., Breland, H., Grandy, J., and Rock, D. (1980). Using longitudinal data to estimate reliability in the presence of correlated errors of measurement. *Educational and Psychological Measurement*, 40(1), 19-29.

Yepes-Bayara, M. (1996, April). *A Cognitive Study Based on the National Assessment of Educational Progress (NAEP) Science Assessment*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.