A Conceptual Framework for the Valid and Comparable Measurement of All Students: An Outline

Rebecca J. Kopriva
April, 1999

The intent of the *Conceptual framework for the Valid and Comparable Measurement of All Students* is to begin to provide a philosophical and methodological basis for the comparable inclusion of all students in testing programs in such a way that, within the constraints of the item types in the test, the scores from the assessment will be reasonably accurate estimations of what diverse students know and can do. This paper extends an earlier paper of the same title (Kopriva, 1996), providing further elaboration, clarification, examples, and a section focusing on research priorities. "Reasonably accurate estimations" about what all students know and can do regarding specified constructs are defined here as scores which have both high validity and high reliability. Currently, it appears that we measure what some students know better, or more accurately, than other students.

I. THE FRAMEWORK

This section will be to present the structure of the framework, define the formal model which underpins it, and discuss some of the implications inherent in why the framework in conceptualized in this way.

Framework Structure

- It has been well documented that students learn differently. They approach new subject matter, process, reason, and store information differently. To teach effectively, educators are regularly encouraged to use a multi-modal approach in their classrooms. Likewise, students bring different strengths and weaknesses to the task of accessing item and test requirements, processing information, solving problems, and articulating responses.

  Since students come to the assessment situations with different strengths and weaknesses, this argues for some flexibility in how evaluations are done. It is tempting to assume all students access test requirements and/or respond to them in the same way. However this is simply not the case. Therefore, it is

necessary to think through how to provide more optimal conditions for all students while maintaining the ability to produce reliable, generalizable, and comparable results.

- It seems reasonable that measurement conditions allow for some variation in selected conditions to more consistently reflect how students access item and test requests and articulate knowledge. Specifically, the proposal to be considered allows items to be presented and administered via a limited, empirically grounded, number of different options, and allows students to demonstrate what know in an empirically grounded set of ways which are meeting their needs, as well as ours. In other words, it is proposed that assessment conditions be allowed to vary within a standard set of empirically supported options, but that they constructs and criteria against which student performance is judged be held constant. Holding constant the accuracy of the information obtained from students becomes, then, the primary anchor which undergirds the comparability of scores.

  To do this, we would move from a single set of measurement conditions (in terms of presentation of items, administration, and response options) to a three dimensional array of conditions which reflect one or more options along the three assessment dimensions. The array is not infinite; in fact, it should be rather constrictive—based on what has been found to be supportable by research and best practice. The parameters need to be defined carefully and specifically so the results are legally supportable. Overall, the array allows for standardization with flexibility, and provides the most optimal set of conditions possible for each student who participates in the assessment.

- There is precedence. The use of rubrics in large scale testing has demonstrated the viability of standardizing constructs and evaluation processes, rather than specific responses, as a way of achieving comparable results. Computer adaptive testing illustrates that students can take different combinations of items and receive parallel scores.

- The current modus operandi in assessment assumes constraints leading to a standard set of conditions for all students, under rationales associated with specifically so the results are legally supportable. Overall, the array allows

the robust comparable and generalizable interpretation of scores. This is not an argument against standardization, but rather an argument that asks psychometricians to rethink what needs to be standardized. *We need to make a choice. Do we want to hold students to one set of conditions or processes, know that because they approach these conditions differently, their understanding of the requirements and their demonstration of mastery DOES NOT give us information across students which is nearly comparable in quality? Or, do we allow some conditions to vary and hold constant our expectations over all students about the quality and accuracy of the information we obtain from our measurements?*

- The answer depends on the inferences to be draw. Any set of measurement conditions favors some students over others. Traditionally, this set of conditions was based on a theory of learning which suggested that students learn in similar ways. Cognitive research has now clearly identified multiple pathways to learning, suggesting the need for multiple or varying sets of conditions to increase the accuracy and precision of measurement when we are evaluating mastery of specified academic constructs. Currently, psychometrics does not know how to obtain generalizable information from students under a varying set of conditions where the quality of inferences are more comparable from student to student.

  The key lies in holding constant the construct inferences to be drawn, rather than holding constant how the information from which we draw the inferences was obtained.

- It stands to reason that, if scores over students are collected using this approach, they should be treated as comparable in decision making and in aggregated and disaggregated analyses. It is true that by providing options additional *potential* sources of error are introduced. However, the options are allowing each student to consistently minimize the distortions between what they know and their scores. Consistent minimum distortions over students, albeit from different sources of error for different students, stabilizes the inferences which can be drawn. Stable inferences suggest that the scores should be treated as comparable, and, in fact, present a definition of comparability which is superior and, in many ways, more stringent, than the ones now used.

Discussion

- We need to determine whether the *processes* of presenting, administering, responding to the construct content support or block a diverse student body from correctly accessing the item/test requirements and demonstrating what they know.

  To do this we need to be clear about specifically what is being measured. We need to be very clear about both the broader constructs we are expecting to measure in the assessment, as well as each construct element that each item is intended to measure. Other than procedures such as those which evaluate differential item functioning, we have not typically been "accountable" for validity at the item level. Further, we have been lax in developing procedures and cross-procedure systemic frameworks which make sure we are measuring what we expect throughout the item and test development, implementation, analysis, and reporting phases. It will be impossible to correctly implement the framework presented in this document if we are not clear about what we are attempting to ascertain. If we aren't excruciatingly cognizant about what we are aiming for, we are not going to effectively correct distortions in our measurements.

- Across subjects, the content of what we want to measure is typically captured in assessments through the types of items and the types of evaluation situations. Item types include multiple choice, short or extended response items, performance task specifications, protocol questions, and portfolio expectations. Evaluation situations include the tradition on-demand testing situation, curriculum embedded testing, observation evaluation, and portfolio collections. Both the item and situation types permit test developers to measure different aspects of mastery in a given subject, where evaluation requirement vary in such ways as depth, scaffolding, and complexity.

  Many of the recent advances in large scale educational assessment have been in expanding our knowledge about and use of a variety of different item and situation types to better measure the academic mastery of students across subject areas. These advances have limited a major source of invalidity as defined by Messick (1989), construct under-representation of a measure's true

score. This occurs whenever knowledge and abilities which are part of a construct are not reflected in the items or tests. However, this same kind of effort has not been made to determine if these diverse types of items are giving us comparable information for all students.

- The framework structure presented above suggests that the processes of the assessment which facilitate the measurement of the construct should also be an integral part of determining validity. High technical rigor shouldn't only be ensuring the viability of *what* is generally measured. It should also provide assurance that *how* the measurement is obtained retains the integrity for each student that what we are expecting to measure is what we are actually measuring.

- Experience in practice has suggested that upwards of 98% of the student population can be measured with essentially the same instrument system if we think about minimizing elements which significantly impair our ability to measure what we are trying to measure. The important point is to remember that getting good information from students is the highest priority. It makes sense to come from this perspective as we develop our instruments, rather than ignoring differences, try to cut and paste after the fact, or by only building systems to effectively measure some students and not others.

- Ensuring that measurement produce valid inferences for everyone who takes the assessment is an essential part of the responsibilities of measurement experts and test developers. This is not the responsibility of the consumer. To date, ensuring the same quality of inferences for some students as we do for others has not been a priority.

Implications

- Treating scores obtained from this type of framework as comparable has important implications for students, the evaluation of school programs with heterogeneous student populations within programs, as well as the evaluation of programs with diverse populations. Clearly defined parameters need to be identified and empirical evidence collected which support the conditions under which this can and should occur.

- In addition to better accommodating students who have identified special needs, it seems reasonable that this perspective might also accommodate a variety of students whose learning, processing, and/or response strengths are divergent from the mainstream. This may attend to potentially problematic legal concerns which deal with the issue of unfairness for those students who might do better with a broader range of format presentation, administration, and/or response options which have not been traditionally allowed.

- It is important to guard against allowing unfair advantage for any students. Thus, "unfair advantage" needs to be defined, and lines which cannot be crossed need to be identified and explored. One such line would be if the varying of conditions results in construct central enhancements for the student rather than minimizing construct irrelevant distortions. Understand this distinction is crucial.

- It is also important to define who selects the conditions to be varied for each student, once the array of possibilities is set for a given testing system. Certainly, this is influenced by the assessment opportunities the students have encountered in their classrooms, the pros and cons of student and teacher choice, et cetera.

- This perspective is based on the belief that the educational measurement field needs to understand more specifically and define more precisely what is actually being measured in our assessments. This begins with defining and/or specifying what each of our items is measuring for each student who takes the assessments, and includes placing a high priority on continuing to develop technically sound procedures to reliably evaluate validity for all who

are tested. This framework suggests that the best procedures to evaluate validity will be systemic, multi-modal (different approaches based on different student strengths), and multi-process (that is, validity will be evaluated over several points throughout the test development and implementation phases).

- In developing more in inclusive assessments, a recognition of diverse access, processing and response needs can and should be built into instruments from the ground up, ensuring that validity and comparability requirements are met at several points throughout the development process. This would replace the current trend which attempts to retrofit assessments ad hoc. Model systems should be conceptualized which specify which combinations of conditions must be met to ensure both technical rigor and accessibility. These systems should include illustrations for different ways to meet these conditions.

The Model

- Accuracy problems appear to occur when construct irrelevant variance *systematically* confounds evaluation of what some students know (Messick, 1989). This type of error distorts inferences obtained from observed scores which have been typically defined to be a function of true scores and random error. The definition of scores under the expanded conceptual framework which includes systematic error is that the observed score is a combination of the construct ("true") score plus systematic, construct irrelevant error plus random error. To formalize this, let y represent the observed score, c represent a desired construct score, s_i represent systematic measurement error and e represent random measurement error.

(1) $$y = c + s_1 + s_2 + \ldots + s_k + e$$

- Current terminology in the new, AERA/APA/NCME draft measurement standards refers to the variables which contribute to construct irrelevant variance as construct irrelevant components (Haertal,1998). Wiley (1997) interprets Messick's definition of relevant variance to be variance derived from construct components which are directly part of the specified construct.

Because of this, he suggests that Messick would classify certain construct components which are not part of the construct, per se, but which are almost always required for used of the components directly defining the construct, as contributing to the irrelevant variance. These components then, are construct irrelevant components.

Kopriva (1997) distinguishes between central construct irrelevant components and non-central construct irrelevant components, where the central components are those which are not part of the construct, but are usually prerequisites to certain knowledge or used in the demonstration of that knowledge, e.g., to solve an algebra problem in mathematics usually requires using arithmetic as well as algebraic skills.[1]

Non-central construct irrelevant components are skills or knowledge which are needed to answer the question but are not materially related to the construct. One example would be the linguistic load paper-and-pencil tests carry in attempting to measure students knowledge in subject areas other than language arts. For linguistically challenged students (for instance those who are limited English proficient or have a disability in reading or writing), this component confounds our capability to measure subject matter knowledge in such areas as science or mathematics via a paper-and-pencil test. Another example would be the need for prior knowledge to answer a question correctly. For example, information about the physics of falling bodies may be needed to solve a particular quadratic equation problem which used this situation to provide a context, though this knowledge would not be required in another problem that is an equally valid measure of the quadratic equation construct.

To formalize the model which includes central and non central construct irrelevant components, equation 1 can be expanded, that is, s_i can be defined as a_i, representing the effect of one of several central irrelevant components,

[1] While these central abilities are defined here to contribute or irrelevant variance, (as defined by Messick and interpreted by Wiley), they are actually relevant to, although not part of, the intended construct. It would also be possible, and perhaps more correct, to suggest that the abilities contributing to relevant variance could be divided into two types—those which are part of the construct and those which are relevant to, but not part of, the construct. In this case irrelevant variance would be derived from only non-central ancillary abilities.

and b_i , representing the effect of one of several non-central irrelevant components.

(2) $$y = c + a_1 + a_2 + \ldots + a_k + b_1 + b_2 + \ldots + b_l + e$$
$$= \Sigma_i a_i + \Sigma_i b_i + e$$

Note that the a_i and the b_i can represent either main effects or interactions. In the latter case, the effects of c and a_i or b_i on y are interdependent.

As all abilities-both construct relevant and irrelevant components are generally correlated, this implies that observed score variance contains covariance as well as variance terms.

(3) $$\sigma_y^2 = \sigma_c^2 + \Sigma_i \sigma_{ai}^2 + \Sigma_i \sigma_{bi}^2 + \Sigma_i \sigma_{aic} + \Sigma_i \sigma_{bic} + \Sigma_{ii'} \sigma_{aiai'}$$
$$+ \Sigma_{ii'} \sigma_{bibi} + \Sigma_{ii'} \sigma_{aibi'} + \sigma_e^2$$

As implied above, the central and non-central construct irrelevant components identified in equation 2 produce what Messick's "construct irrelevant variance". Note that σ_{ai}^2 and σ_{bi}^2 are variances and σ_{aic}, σ_{bic}, $\sigma_{aiai'}$, $\sigma_{bibi'}$ and $\sigma_{aibi'}$ are covariances.

One goal inherent in these equations is that y should represent c and a_i to the largest extent possible. This means maximizing the relationship between y and $c + a_i$ by minimizing the distorting effects of the b_i on y. It also seems important to differentiate between c and a_i in the model, and in explanations of how the construct is interpreted. The a_i, or central irrelevant components, become the essential building blocks for operationally defining and restricting the construct, and so it is important to identify the a_i, separating them from the conceptions of the construct. Therefore, while c and a_i are not differentiated in the observed score, or y, the distinction should be made in constraining the inference of y.

- This model can also be interpreted as a structural model where various latent causal factors (the c, a_i, b_i, and e) impact the observed score (y).

In Figure 1, the ovals represent the latent causal factors which can contribute to the observed score, show in the square boxes. Error, or e, also affects the observed score. The arrows among the latent factors reflect their interdependence. The arrows from error and the latent factors to the observed score records the impact of that variable on the score. In each case, the impact of the variable on the score can range from -1 to +1, with values towards 0 reflecting little impact, and absolute values towards 1 reflecting greater impact.

As noted above, the goal of the framework presented here attempts to present an observed score which reflects the student's ability on the latent construct (*c),* plus central irrelevant components (a_i), with a minimum amount of distortion. This means maximizing the impact of c and the a_i on y . The relationship between the construct (c) and the score is defined to be 1, as c constitutes the primary focus of the measurement. The relationships between the a_i and y are high, assuming maximization.

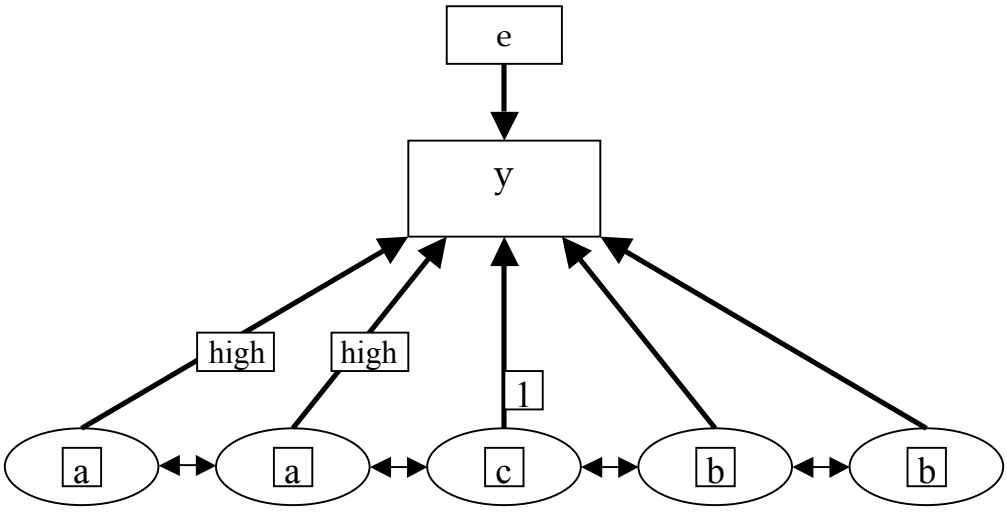

Figure 1

- In fact, however, the observed score (y) is always a function of the conditions under which it is measured. Therefore, y is actually $y(x_k)$. $y(x_k)$ is defined as the observed score under a set of measurement conditions, x, of which each condition of a set can be labeled as x_k. The set of conditions include presentation, administration, and response conditions under which the

content is introduced to the student of the information about his or her knowledge is collected. The observe score, depicted in this fashion, is illustrate in Figure 2.

- We know that, due to individual strengths and challenges, non-central construct irrelevant components (the b_i) affect students differently. Further, there appear to be systematic patterns of particular components which affect subgroups of students differently. The focus of this framework is on the a subset of b_i s which interact with the conditions of measurement. These b_i can be defined as b_k where k is one particular condition of a set of conditions of measurement (x_k).

As can be seen in figure 2, y is observed under two conditions of measurement, (x_1) and (x_2). Two b_ks, b_1, and b_2, in addition to the a_i, c, and other b_i, impact $y(x_k)$. The b_k which corresponds to the same measurement condition (x_k) impacts $y(x_k)$ highly, where the other b_k does not .

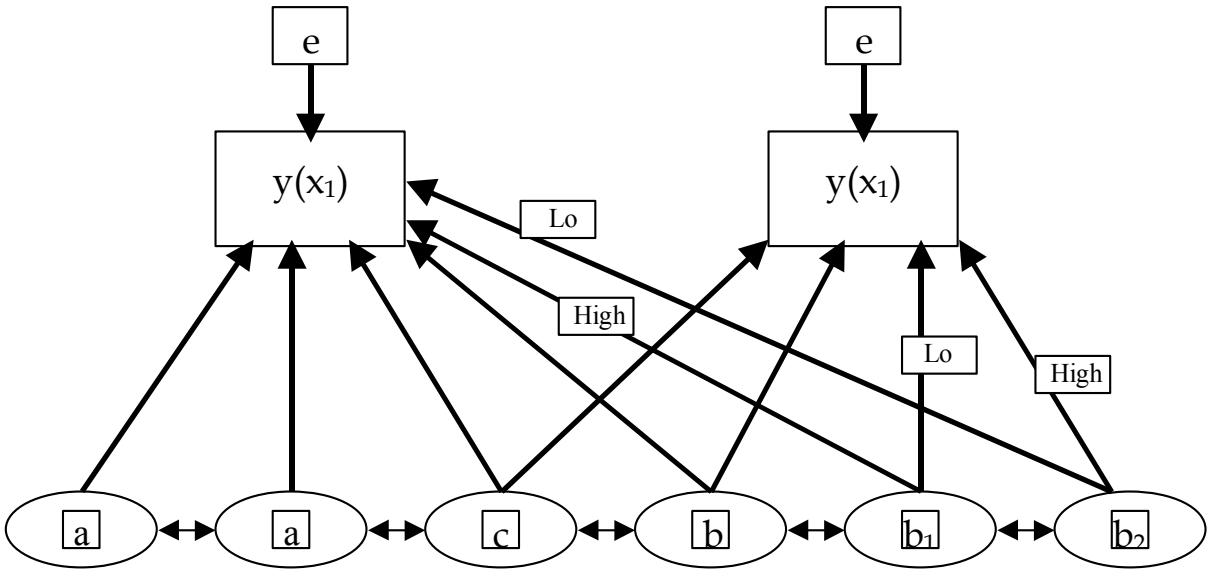

Figure 2

- The same b_k s which correspond to the $x_k s$ in a specific measurement situation are the latent factors which can highly distort the $y(x_k)$. Within students, the levels of the b_k can range from low to high. In other words, b_k could be student reading ability, and a student can read poorly (low level of b_k) or well (a high level of b_k). $Y(x_k)$ is minimally distorted when the level of the corresponding b_k is high for a given student. On the other hand, the distortion will be high when the impact of the b_k on $y(x_k)$ is high—that is, the k's correspond, and the level of this factor for the student is low.

For example, two LEP students are expected to take part in a mathematics assessment. Student 1 has a high degree or oral fluency and literacy in his home language, Spanish, and a low degree of proficiency in English. Student 2 is not literate in Spanish or English, but he is orally proficient in Spanish. For low distortion to occur for Student 1, there must be a match between his high levels of proficiency in Spanish and the same measurement conditions, such as a mathematics paper-and-pencil test in Spanish, or a mathematics test given orally in Spanish. Likewise, Student 2 would need to match his high degree of oral fluency in Spanish with the measurement condition of oral administration and response (if necessary) in Spanish for there to be a low distortion is his observed score on the mathematics assessment. In both cases, if the measurement conditions do match b_k latent component factors where there is a low level of proficiency for each student (e.g. in reading or writing English), then high distortion would occur. High distortion would also occur for Student 2 if he is administered the assessment as a Spanish paper-and-pencil test.

- As was discussed above, the set of measurement conditions under which information about student mastery is collected occurs at one or more points in the presentation of the assessment, the administration, and the response opportunities associated with the data collection. The examples above focused on only two aspects of the set, that is on the language presentation condition and oral vs. paper and pencil administration. In reality, for low distortion to be achieved in the observed score, all measurement conditions would need to match a student's high level of ability in each of the corresponding b_k. To the extent that the measurement conditions set chosen for the student corresponds to a mix of latent components where the student

has high and low levels of proficiency, the distortion in the $y(x_k)$ will heighten accordingly.

Is this approach psychometrically and legally defensible? Is it worth the expense? Can we afford not to do it? Are we ready to face the consequences of the constraints we have traditionally placed on ourselves in the name of standardization, given the advances in cognitive psychology about how children learn, process and respond? If the answers to these questions are yes in some cases and no in others, what are those cases and what are the parameters which define them, psychometrically, legally, and practically?

It will be important to identify and prioritize what sorts of empirical information need to be collected to support or dismiss this conceptual perspective, and under what circumstances. Phil Daro (August 1996) suggested that, even given recent advances, the performance of about 1/3 of students appears to not be validly assessed. If he is even close to correct, it seems that serious thought needs to be given to continuing to investigate how scores are obtained and how they might be treated relative to one another. One approach is presented here.

II. EXAMPLES

In this section a mathematics item will be presented, and a preliminary analysis of what it appears to be measuring will be outlined. Subsequently, two alternatives to the item will be identified and discussed as well. The alternatives will be items where the same mathematical construct elements appear to be measured as those in the original items, but the item requirements of additional, non-central latent components irrelevant to the constructs will change. In the language of the model discussed above, the measurement of the central "c" and "a_i" will remain constant across each of the items and their alternatives, while the "b_i" will differ.

The purpose of this section is provide examples of how items might be allowed to differ, while still measuring the same important subject matter construct elements. It is assumed that, if the same construct elements are being measured across items for each form in a given assessment, then the same constructs are being measured across forms.

Base Item

Tom knows that 50% of the students from his high school are accepted at the local community college, and 50% are accepted at the state university—so, he reasons that he has a 100% chance of being accepted at one place or the other is he correct? Explain your answer.

A preliminary analysis of what the item is measuring follows, that is, the construct (c), the central construct irrelevant latent components (a_i) and the non-central construct irrelevant latent components (b_i).

Preliminary Analysis

c disjoint sets

c recognizes fallacy

a communicating mathematical ideas (using diagrams, graphs, etc.)

a 50% means one half of set
a 100% means the whole set
b universal knowledge about college admissions
b knowledge of high school to college culture
b ability to read in English—rate, fluency
b ability to understand requirements without additional clues, resources (e.g. visual, kinesthetic)
b U.S. language, vocabulary
b ability to write
b understanding of desire response mode
b ability to explain or communicate
b discourse practice
b test behaviors

Alternative #1

50% of the students like apples and 50% of the students like oranges. Mary thinks that Tom, a student in the class, likes one fruit or the other. Is she correct? Explain your answer.

Preliminary Analysis

c disjoint sets
c recognizes fallacy

a communicating mathematical ideas (using diagrams, graph, etc.)
a 50% means half of whole set

b universal knowledge about fruit
b ability to read, but less linguistic load than base item—less words (34 to 49, more readable words, simpler sentence structure, male and female in item
b ability to write
b ability to understand requirements without additional clues, resources (e.g. visual, kinesthetic)
b understanding of desired response mode
b ability to explain or communicate
b U.S. language

b discourse practices
b test behaviors

This alternative changes the context of the item to a context which is more widely known (fruit), and there are changes in the overall readability of the item. The clarification of the desired response mode, paper-and-pencil format, in English, have not changed.

Alternative #2

50% of the students like the color red and 50% of the students like the color blue. Mary thinks that Tom, a student in the class, likes one color or the other. Is she correct? Explain your answer.

(Include visual with two circles, half lightly shaded in each. Students are invited to color in circles if they want.)

Preliminary Analysis

c disjoint sets
c recognizes fallacy

a communicating mathematical ideas (using diagrams, graphs, etc.)
a 50% means half of whole set

b universal knowledge about colors
b ability to read, but less linguistic load than base item—less words, more readable words, simpler sentence structure, male and female in item, and more alternatives to access requirements. (visual and kinesthetic)
b ability to write
b understanding of desired response mode
b ability to explain or communicate
b U.S. language
b discourse practices
b test behaviors

This alternative introduces a visual (the circles), and an optional kinesthetic activity (coloring the circles if they want). This presents more stimulation, some more words (invitation to color circles), but allows student to access the item requirements through means other than words.

III. EMPIRICAL EVIDENCE

There appear to be three central issues or questions, and associated supporting concerns, which must be addressed in a research agenda designed to investigate the main tenets of the framework. Certainly, this agenda will build on recent and current research which is being conducted in a variety of arenas. These include research in accommodations and plain language editing for special needs populations, in studies which focus on how children process and articulate information in various settings, and investigations dealing with the determination of what is being measured under different circumstances.

First, it seems essential that a systematic approach must be identified which defines and confirms what subject matter construct and construct elements are actually being measured at the test and item levels, respectively. Our approach has been to focus on a face matching of test specifications to a narrative of the construct domain or content standards, and a face matching of items to test specification cells. Careful thought has not been give to what else we might be measuring. We also have not systematically investigated if, in fact, we are measuring what we intend. Further, we have not determined whether or not we are measuring what we intend for each student who takes the test. Work has been done in several projects which can guide how this might be accomplished, including several in the field of cognitive psychology (Greeno, 1998), the test development inclusion guidelines for measuring the mastery of LEP student (Kopriva, 1998), the assessment development efforts in the New Standard Project (Shannon, 1996; Tanner, 1997), and preparations for the Voluntary National Test (Baldwin, 1998). However, to date no systemic effort has modeled and investigated the set of processes which must be accomplished to determine what we are measuring.

The second primary question focuses on the flexibility premise of the framework. What types of options can be employed and when, in order to provide assessments where students are provided a fairer opportunity to access testing requirements and demonstrate what they know (while the construct continue to remain constant over forms)? Once we know how to determine and confirm what we want to be measuring, we can design studies which can begin to determine the range of options which might be tolerable, and for which students.

Two sub-questions are also central. Are the options collected under the varying conditions giving us more accurate information, for whom and in what situations? When is the information collected under varying conditions assumed to be comparable, and when isn't it?

Evaluating the research data in terms of increased accuracy provides understanding about the effectiveness of the options. Determining comparability (and aggregatability) will bear on the usefulness of the flexible testing systems. Certainly, the research will need to call upon recent work about how to determine what students actually know so that this evaluation can be used to benchmark the effectiveness of the option-based scores. It will also need to build on work which extends the concepts of comparability beyond traditional notions of building parallel tests. These investigations should begin to define parameters which can guide practical decisions as well as clarify directions for further research.

The third issue revolves around the impact of replacing or supplementing traditional types of tests with this broader conceptualization of testing. Considerations include the impact on the development of assessment. For example, what guidelines need to be identified in test specifications, how do these impact the collection of confirmatory information in pilot and field test, and when and how are cutscores and/or normative data pooled across options? Deliberations should also impact the practical issues associated with test implementation (e.g. who decides what options and for which students), provide a lens to refine how we interpret results, and consider the impact of types of testing on test use, particularly in high stakes accountability climates. Political ramifications and guidance about how to communicate testing modifications to measurement colleagues, the broader education community, parents and the general public are also important. While the focus of the agenda is to collect empirical information, how the results should be integrated into these areas will need to be addressed as well.