Evidentiary Logic in the Assessment of Diverse Learners:

Valid Assessment of English language Learners (VAELL) Project

Rebecca Kopriva
David Wiley
Phoebe Winter
University of Maryland College Park

April, 2003

Paper presented at the Annual Conference of the National Council of Measurement in Education, April 2003, Chicago, Illinois.

I. Introduction

In his paper on evidence and inference in educational measurement, Mislevy (1994) makes the point educational assessments provide data in a variety of forms. He then goes on to say “these data become evidence only with respect to conjectures about students and their work – conjectures constructed around notions of the character and acquisition of knowledge and skill, and shaped by the purpose of the assessment and the nature of the inference required” (pg. 441). In order to capture the nature of the ‘character and acquisition of knowledge and skill’, perspectives on cognition and learning must be considered. Then, what is known about cognition can be applied to the measurement of student achievement.

The VAELL project is addressing two purposes. First, we are interested in investigating evidentiary chains that will lead us to new models of test performance conditioned upon the complex interactions between how targeted knowledge and access jointly impact student performance. These interactions occur when students with certain levels and types of access interface with items that are designed to communicate meaning through the use of specified conditions. It is anticipated that the new models will inform if and how students can take tests under varied conditions in such a way that the scores across conditions can be seen as comparable.

Second, the VAELL project is researching how to develop large-scale evidence-based assessment options that successfully meet both the access and score comparability requirements for students with different needs and challenges.

Both sets of investigations are focused on identifying suitable access for English language learners as well as other students with language, literacy, or cultural challenges. Presented below is a brief description of the project and the process and progress to date regarding each of the two purposes of the project.

II. Overview and Rationale of the Project

In our view, test validity issues center on the notions of target constructs and construct irrelevant factors (Messick, 1989). A target construct embodies the measurement intent of the test. Construct irrelevant factors are those things that the test measures that are unintended and not conceptually related to the measurement intent.

Modern large scale test-making attempts to minimize construct-irrelevant variance. Generally this means that publishers consider and attempt to minimize non-relevant deficits students may have that will get in the way of measuring the construct that is supposed to be reflected in the items by standardizing conditions under which students are assessed. However, at times, these deficits are not minimized for specific groups of students. In these cases, systematic errors due to non-relevant factors accrue over items to create measurement distortions. For example, linguistically complex mathematics items may cause difficulties for students with low levels of either reading ability or English language acquisition.

Items may not only embody barriers to a student's item access, but may also be constructed so as to permit pathways to the item's content. E.g., for poor readers with appropriate visual facility, a linguistically complex item incorporating a graphic may give an alternative access to item content.

In formal terms access then can be specified as an interaction between item features and person characteristics that either permits or inhibits the student's access to the actual content of the item.

Most research about access, however, ignores the complex nature of the relationship between access factors and the demonstration of targeted knowledge. This is especially true for English language learners, other language and cultural minority students, and those with sizable literacy and attention-based challenges (Kopriva and Lowrey, 1994; Abedi et al., 1999; Shaw, 1998) Because of the nature of how today's standardized tests request and communicate information, the language, cultural and contextual challenges of English language learners and others form a web of small blockages that pervade tests and test items, and are impervious to easy post facto accommodation attempts. While individual challenges or mismatches are small, the sum total of the blockages in and over items present a sizable distortion in the data when the academic proficiency of many English language learners is measured in traditional ways. Many, including Solano-Flores and Trumball (2003) and Kopriva (1999), argue that existing approaches do not ensure equitable and valid outcomes.

Clearly, the substantial work in large scale measurement can and should not be set aside. Rather it is suggested that a proper range of standardized and empirically-based item and administration options be included in large scale testing systems in a format that would allow students to demonstrate what they know and still allow for the aggregate-ability and comparability of scores over options.

Such an options inclusive testing system would ground the comparability of the scores in the development of item and test options that are seen as synonymous with respect to their targeted knowledge function. In other words, assessment conditions could vary in prescribed ways but *the constructs and criteria against which student performance is judged would be held constant*. While this is a different way of thinking about validity and comparability, it is not without precedent. One example are scoring rubrics used in large scale testing which, with appropriate technical and oversight supports, standardize the constructs and evaluation processes while allowing specific responses to vary from student to student. Another example is with computer adaptive tests, which provide different combinations of items to each student, yet yield measures of ability that are comparable. Certainly, such a system would need to address how and when additional sources of error due to options would offset errors occurring from decreased validity of score inferences for specific students.

The focus of the work in VAELL will be on identifying and investigating construct irrelevant factors that are particularly problematic for English language learners and other student with similar challenges. Overall, this suggests that the types of options would be primarily options in how items are presented to students, including what is available within the presentation to help students apprehend the meaning of the item, formulate the problem and be able to explain their answer appropriately. Administration options will also be investigated but are overall somewhat less prominent for this population.

II. Model Conceptualization to Date

In the field of assessment, models are generally fixed and inferences are formed that are conditional upon those models. In this project we seek to generate evidence about processes and in doing so create evidentiary chains that lead us to new models. Though we begin with models that are accepted within the psychometric community, they are meant to be a starting point in this process. In addition, using established models provides a vocabulary with which to discuss the measurement of proficiency.

In 1980, the multicomponent latent trait model (MLTM) was proposed (Whitely, 1980) to model the components underlying item response processes within an IRT framework. This model denotes the probability of success for a person on an individual item as the product of success probabilities for each of the underlying components. This can be written as follows:

$$P(X_{isT} = 1 \mid \underline{\theta}_s, \underline{\beta}_i) = \prod\nolimits_m \frac{\exp(\theta_{sm} - \beta_{im})}{1 + \exp(\theta_{sm} - \beta_{im})} \qquad (1)$$

where $\underline{\theta}_s =$ the trait levels of person s on the M components

$\underline{\beta}_i =$ the difficulty of item i on the M components

$\theta_{sm} =$ the trait level of person s on component m

$\beta_{im} =$ the difficulty of item i on component m

Due to the multiplicative, rather than additive, nature of this conjunctive non-compensatory model a deficit in proficiency on any of the components will lead to a smaller probability of getting the item correct. In this model, unless the person parameter is significantly larger than the item parameter for all m components, the probability of a correct response to the item will be relatively low.

Building upon the MLTM, we propose that the probability of a correct response to an item can be represented as the product of the construct relevant components of an item and other components, such as those of access, that would be considered construct irrelevant or ancillary to the domain of interest. Continuing with the Rasch framework of the MLTM, the probability of success on an item using this model may be expressed as:

$$P(X_{isT}=1|\underline{\theta}_s,\underline{\beta}_i)=\frac{\exp(\theta_{sa}-\beta_{ia})}{1+\exp(\theta_{sa}-\beta_{ia})}\cdot\frac{\exp(\theta_{sc}-\beta_{ic})}{1+\exp(\theta_{sc}-\beta_{ic})} \qquad (2)$$

where the subscript a in the equation refers to the ancillary components and the subscript c refers to the construct relevant components.

There is another class of items in which there is a compensatory nature to the measurement components. McKinley and Reckase (1982) proposed the multidimensional Rasch model shown below which models this phenomenon.[1]

$$P(X_{is}=1|\underline{\theta}_s,\delta_i)=\frac{\exp(\sum_m\theta_{sm}-\beta_i)}{1+\exp(\sum_m\theta_{sm}-\beta_i)} \qquad (3)$$

where $\theta_{sm}=$ the trait level of person s on dimension m

$\beta_i=$ the difficulty of item i

In this model, the undimensional trait level is replaced by an equally weighted sum of the composite traits. In this configuration a lack of one trait can be made up for by the presence of another.

This compensatory model would seem to be particularly useful as a starting point for understanding how access factors interact in an item for a particular student. For

[1] The notation of McKinley and Reckase includes an easiness parameter rather than the difficulty parameter shown here. Their notation has been changed to be consistent with the other notation used in this paper.

example, a student that with a low level of reading literacy may be able to compensate for that deficit if the item provides a visual that makes the meaning of the question clear and the student can decode that visual in a meaningful way.

In this project, we propose using the compensatory approach to model access, while also attempting to capture the conjunctive relationship between access and construct status for both student and item factors. For instance, we might build the compensatory element into the conjunctive model given in equation (1) yields the following:

$$P(X_{isT} = 1 | \underline{\theta}_s, \underline{\beta}_i) = \frac{\exp(\sum_{m_a} \theta_{sm_a} - \beta_{ia})}{1 + \exp(\sum_{m_a} \theta_{sma} - \beta_{ia})} \cdot \frac{\exp(\theta_{sc} - \beta_{ic})}{1 + \exp(\theta_{sc} - \beta_{ic})} \quad (4).$$

However, we are concerned that the relationship isn’t consistently as straightforward as equation 4 would suggest. Rather, we believe we will find in our research that under specific cases or types of item-student and/or access-construct interactions, the relationships will reflect a weighted and more nuanced approach to student performance. For example, how items with the same vectors of item factor levels evoke required or permissible student factor levels for successful performance will often differ depending on the joint status of the factors and other characteristics of the items. Further we expect that these relationships can be explained and be model-able. Therefore, for now, we prefer to identify the multidimensional access process that captures the compensatory and non-compensatory factors jointly as a more general interactive function,

$$P(X=1| \Theta, A) = F(A. \Theta),$$

where A has been used to denote access, the probability of success on an item is a function of latent ability, the relevant item characteristics, as well as construct-irrelevant person and item characteristics. Here, $0 \leq A \leq 1$. $A = 1$ implies full or complete access, $A = 0$ implies no access, and $0 < A < 1$ implies intermediate or partial access.

Modeling the item to student access matches An important part of modeling the access-targeted knowledge relationships is to model how the student and item access factors interact. To date we are finding that item to student access matches seem to follow a series of procedures and rules that include

1) identification of the access factors evoked by particular items across students to form a access profile vector by item
2) identification of interactive rules within the item profile vector for specified factors across students with different student profiles
3) identification of access factors in student profiles across items to form a vector profile by student
4) weighting of access factors within student vectors to provide prioritized matching expectations that differ by student.

Identification includes the presence or absence of specific access components in items or in students, that can be matched from student to item, and that quantify threshold amounts of each access component. For example, an item may use reading as an access factor in a mathematics item, and the threshold for the reading level could be identified as high. The students would have a reading factor level identified on the student access profile that would range from low to high.

Interactive item rules explain when the item can be accessed by more than one factor or when the presence of one factor could lower the threshold for another factor.

For students, some factors may influence the profile more heavily and need to be weighted accordingly. For instance the student may need culturally appropriate items, regardless of the visual aspects of an item and whenever the item is at least at a mid-range reading level.

Preliminary work in modeling these matching procedures and rules has suggested that they are model-able and programmable--for students with different access profile vectors and targeted knowledge levels, and for items and item types that measure various types of targeted information and evoke different access needs. In general the approach has been to create an access profile vector for each student that will be used across items, and an item access vector for each item that explains what access factors are evoked in that item. In each case, as noted above, the student and item vectors are "customized" by additional rules of compensatory and conjunctive interaction and weighting that differ over students and items, respectively. Once these profiles have been identified and the rules programmed by student or item, it appears that the matching can be completed electronically. The product is a best match among like items can be found for each student for each identified targeted construct element that a test is intending to measure. A discussion of how “like items” are identified as such and developed is discussed in the section below entitled *Accessible Test Development to Date*.

Additional model components Our initial findings suggest that there appear to be three main points within the problem solving process where access is crucial:

- Apprehension of the task demands
- Solution formulation
- Articulation of the task solution

In order to respond correctly to an item, the test taker must first be able to access and apprehend what the task requires in terms of a valid solution. Depending on the nature of the task, accessing its demands may involve reading text, interpreting graphics or decoding symbols. An obvious example of not being able to access the demands of a task is an item presented in English to a student who cannot read English. After accessing the task on this surface level, the examinee must also be able to apprehend its meaning in terms of the desired outcomes. Students who fully understand the meaning of the task at this deeper level will be able to move to the next cognitive phase: problem solving.

The solution formulation stage of the task solution process can be broken down into two components: problem representation and solution strategies. In order to respond correctly to a task, the respondent must first create cognitive schema to represent the problem in away that can serve as the basis of a solution. Access to construct-relevant schema, as well as, student access to the manner in which the task is presented affect these problem representations. There may be multiple appropriate representations to a single problem that correspond to alternative solution strategies that might be employed. In order to carry out a solution strategy the examinee must have access to the cognitive processes necessary to implement a solution and then use these processes to actually achieve a solution. It is important to note that a test taker may formulate a solution strategy compatible with the problem representation established but still not be able to carry it out. For example, a student may know what the perimeter of a figure is and understand that to arrive at an answer the figure must be measured in some way. If the student does not understand that an approximate measuring tool can be constructed from the materials on hand, he or she cannot complete the solution process.

The final step in solving a task involves the articulation of the solution. Though this may seem little more than simply stating the answer, this process is actually fairly complex. First, the student must apprehend the solution, which involves combining part of accessing the original task demands with the solution articulation. Then it is necessary to develop the required response by accessing the cognitive processes needed and using these processes to achieve the response. Finally, solution articulation involves the ability to represent the response. Take for example, the situation in which a pictorial response is most appropriate for a constructed response item, but the examinee is inhibited from drawing due to a lack of skill. In this case the student will be forced to attempt to verbally explain something that might have taken a relatively short time to draw.

It is expected that access for each point of the problem solving process will ultimately be factored into the item vectors and possibly the student profiles. Because only a restricted set of access factors will be evoked at any one point in the process, it is expected that this refinement will add an additional set of rules to the established vectors.

IV. Accessible Test Development to Date

The other primary focus of the VAELL project is to research and develop procedures and model products that are designed to meet the needs of large-scale testing and that address both the access and score comparability requirements explained above. To date this has included

- the identification of an item development approach that meets both needs
- the identification of the development process and set of procedures
- the development of the first set of model items that fit these specifications
- an initial empirical investigation into the validity of this approach.

Item development design and implementation The focus here was to design an approach to item development that would be rigorous for large scale assessment and still meet the needs of diverse types of students. Consistent with the rationale for the project explained above, this meant that the approach should measure the same targeted knowledge over some amount of items, but that the construct irrelevant factors could differ to minimize blocks and permit access. The development would take place in mathematics at grades 3 and 5.

To address this focus, the concept of item clusters was adopted and implemented. The figure below illustrates the cluster concept. All items in a cluster share the same targeted core content and the same cluster specific mathematics, and differ only in the type and degree of ancillary factors they evoke. Core content is defined as mathematics specifications of targeted knowledge that is being measured by an item that embodies it. Cluster specific mathematics is mathematics and/or context specifications that are not targeted but that need to remain consistent across identified items for those items to be seen as variations on the same mathematics ideas. In order for each item to measure the same mathematics construct, the item must measure and utilize both the same targeted and cluster specific knowledge and skills. An example of two items in a cluster would be an item that is written in English and the same one that is written in Spanish.

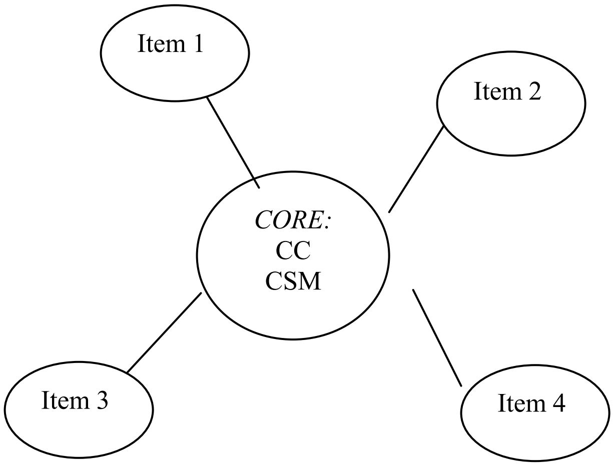

Figure 1. Item Cluster

The Item Cluster concept represents a template approach that is more constrained than many envision when they discuss the use of templates or item shells. However, in constraining the template by including cluster specific as well as core content specifications, the concept minimizes error due to many of the additional sources of variation in items. Such sources of variation would be those that occur across items that are derived from templates whose criteria include only the content objectives of items. This allows for the construction of a set of items that are virtually interchangeable, while also meeting the needs of students with diverse access challenges. It marks an intermediate step between one item and items that share specified general template criteria.

In general, the core of the cluster, consisting of both the targeted knowledge and the cluster specific specifications, would be identified prior to item development and a set of items would then be constructed to each measure the core but minimize specific access factors. In order to be able to validate the viability of this approach for today's state testing needs, VAELL identified released items from testing systems across the country. The core specifications for those items were then defined, and more items were created that conformed to the same core specifications. For each released item, the core specifications became the core of the item cluster, and the released item became one of the items within that cluster. Item access profile vectors described above were completed for the released item and for each new item in each cluster.

Clinical Investigation A large scale empirical study associated with both the cluster approach and the model conceptualization is planned. Currently, a clinical investigation is underway.

Initially, mathematics experts identified algebra and number as the two mathematics strands that would be studied at grades 3 and 5, released test items that conformed to these two domains were identified, and clusters completed for each grade.

Items were typically written by mathematics educators, and educators with expertise in linguistics in K-12 education and English language learners. Completed clusters were informally reviewed by mathematics experts to ensure that the same core was retained. A formal blind review by mathematics experts is planned.

In the clinical trials, the items and clusters, the general model approach discussed above, the model matching process, and the viability of the problem solving access points are being investigated. Approximately 175 students in grades 3 and 5 were selected who represented the full range of SES, and mathematics knowledge. Students in both grades were stratified re. their status as currently receiving ESOL services, former ESOL, or native English speaker, and those with IEP in reading.

To collect information about the student's access factors, teachers responded to a questionnaire. Items were selected as a best match for the students given the questionnaire results, and 4 of those items (the preferred set), along with 4 parallel release type items that measure the same core (the matched set), were administered to the students. Students were then interviewed about their thinking on both sets of items in an attempt to understand if and how access was involved in the student's problem solving. Follow up interviews were also conducted with teachers to collect independent information about the students mathematics knowledge in the targeted areas, teacher instructional and testing strategies that appear to work and not work for each student, and other related information about the student factors.

The clinical trials will be completed mid-May. Analyses will focus on the viability of the cluster approach and the model components, and on how factor vectors might be improved and streamlined for large scale administration.