

Testing for the Future: Addressing the Needs of Low English Proficient Learners Through Using Dynamic Formats and Expanded Item Types

Rebecca Kopriva
University of Wisconsin Madison

James Bauman
Center for Applied Linguistics

The Problem

Two sets of studies from two different states illustrate the reason why improvements need be made on how the field measures the academic knowledge of students with low English proficiency. While advances in accommodations have impacted how well academic content is assessed for more advanced ELLs, the language and cultural challenges of beginners and some intermediate students do not appear to have not been properly mitigated to date.

In Case 1, shown in Table 1 below, a study evaluated the distractor distributions of select multiple choice items from an operational statewide test where item writers were well versed in item principles of universal design. The items were sampled from four subject areas and six grades. The analyses compared scores from each of three treatment groups, ELLs who were proficient at prefunctional, beginner, and intermediate levels (PBI), students with learning disabilities, and hearing impaired students, to the control group of native English speakers without IEPs. Samples were balanced for each group dyad. The assumption was that distractor distributions of students who got the item incorrect should be similar across treatment/control groups if the items are accessible for students in the treatment group. As Table 1 illustrates, the percentage of items significantly different for each of the treatment/control group dyads is substantial. The most profound impact is for the lower English proficient ELLs, where 63% and 71% of the distractor distributions were significantly different from the control group.

Table 1 Analysis of State Operational Test Data: Treatment Groups vs. Control

<i>Subject</i>	<i># Items</i>	<i>Grades 3-5</i>			<i>Grades 6-8</i>			
		<i>% Problems*</i>			<i>% Problems*</i>			
		PBI ELL	LD	Hearing Impaired	PBI ELL	LD	Hearing Impaired	
<i>ELA</i>	6	33	33	50	4	100	25	50
<i>Math</i>	6	67	50	33	16	63	69	63
<i>Science</i>	8	63	50	63	7	86	71	71
<i>Social Studies</i>	4	100	50	50	7	57	29	29
Mean %		63%	46%	50%	71%	56%	56%	

* Chi-square results for treatment group are significantly different from those of control students

Findings from the second case are illustrated in Table 2. The results are from recent criterion validity study which investigated the relationship between ELL scores (3 levels of proficiency) vs. scores of exited and native English speakers on multiple choice and constructed response items from a large-scale mathematics test at grades 3 & 5. The criterion indices that were used were teacher ratings of student's skills in specific mathematics elements covered by the test. Study staff worked with district staff to develop a test which used plain language and various compensatory supports, and this test was given to all students; study staff were also responsible for carrying out the various administration accommodations to ensure quality control across schools. Given the amount of high quality accommodation support, the hypothesis was that the relationship between criterion ratings and scores (as exemplified by regression and discrimination coefficients, R^2 s and Betas) would be similar for most ELLs as compared to the control groups of exited and native English speakers. The goal was to find no significant difference between the group dyads. As Table 2 shows, only the advanced students in grade 5 showed no difference for both types of items. Surprisingly, the constructed response scores of beginners at both grades were similar to the control groups as well. Except for these instances, the R^2 and Beta for the exited/native English speakers vs. ELLs were quite different overall.

Table 2 Criterion Validity Study

		<i>Multiple Choice</i>			<i>Constructed Response</i>		
		B	p	R^2	B	p	R^2
<i>Grade 3</i>	Beg	.41	.29	.02	1.41	.00	.21*
	Int	.87	.00	.05	1.15	.00	.06
	Adv	1.11	.07	.05	1.74	.02	.08
	Exit	2.73	.00	.20	3.68	.00	.20
	Nat E	2.67	.00	.27	3.57	.00	.23
<i>Grade 5</i>	Beg	-.32	.50	.01	1.32	.02	.13*
	Int	.78	.01	.04	1.58	.00	.06
	Adv	1.61	.01	.13*	2.91	.00	.20*
	Exit	1.87	.00	.19	3.43	.00	.26
	Nat E	2.25	.00	.23	3.20	.00	.22

* NO significant difference with exited and native English speaking students

A New Generation of Tests

As more formative testing uses computer platforms, and more states consider shifting from large-scale paper and pencil assessments to online delivery systems, it seems inevitable that academic testing will shift to using computer applications to assess students. As this occurs, large-scale computer-based assessments will undoubtedly begin to take more advantage of the computer's dynamic capabilities as compared to the static nature of today's tests. Already, some states (for instance Minnesota) and projects (for instance science assessment research and development at SRI and WestEd), are

investigating and preparing prototypes of how the dynamic capabilities might be harnessed. Further, in the last 10 years or so, several projects have illustrated how the computer-based approach would be useful in systematically administering a range of accommodations online for ELLs or students with disabilities in conjunction with traditional static item types (for example Tindal, 2006; Kopriva et al., 2007). However, to date, no project has researched how dynamic formats may be used in a large-scale testing system and how they might be especially advantageous for students with the lowest literacy in English and possibly others with language or literacy challenges. Two recently funded projects, ONPAR in science and ONPAR in mathematics, are researching the viability of building computer-based dynamic items which harness animation, simulation, and interactive qualities. The focus of the work is on improving the measurement of academic content for students whose English skills are minimal, and on beginning to solve some of the problems inherent in moving testing for all students onto a dynamic platform. The thinking to date behind this work is the subject of this paper.

Many issues need to be considered before dynamic testing can become a mainstay for large-scale assessment. Besides cost for development and implementation, four considerations immediately come to mind. First, it will be important to understand what aspects of dynamic testing are relevant for improving how the knowledge and skills of students can be measured. As testing transitions to computer delivery, what types of measurement tasks particularly lend themselves to the computer capabilities such as those summarized above and below? On the flip side, when are concepts or skills best measured using static formats? Second, the role of expanded item types needs to be understood. While multiple choice has obviously dominated static testing, this was largely because it was easy to scan and score on a mass scale. Over time, a substantial body of literature has grown up that explains the properties and limits of testing using this item format. However, it is by no means the only forced-choice type of item that can be computer-based, and several kinds of items can be scored algorithmically as students complete their work. It will be important to understand the nature of other item types that interact well with basic computer capabilities.

Third, considering dynamic computer-based testing means understanding the cognitive demands associated with this mode of assessment, and, one would argue, better understanding the cognitive demands associated with traditional testing. The task-irrelevant cognitive demands will be considerably different across static and dynamic environments and it will be important to reconcile when tradeoffs are effective and when they might be problematic. Future work will need to consider when items in different modes can be considered comparable, which would seem to assume similarity of item targets across formats, and perhaps similar levels of compensatory irrelevant-demands within each mode. It is probable that the irrelevant demands, which should be used to facilitate the target requirements in items, rather than acting as barriers to the target demands, would impact different students differently. As this is the case, this interaction needs to be understood as well.

Fourth, as mentioned above, the authors believe that the dynamic and related capabilities of computers can be used to substantively address the challenges of students who have little English language proficiency. It will be essential to understand how the capabilities can mitigate the target-irrelevant testing problems faced by these students, even while static item formats may be suitable for other students. Key for understanding these considerations is knowing how much language is still required in the dynamic environment, and when the role of language can be assumed by other item elements.

Each of these questions will be touched on to some extent in the twin projects. Before discussing the projects more fully, the next section will summarize quickly some of the conceptual underpinnings that provide precedence for this work. These concepts are based in the fields of measurement and linguistics.

Conceptual Underpinnings for Creating Dynamic Contexts and Expanded Item Types

Measurement Underpinnings

- In explaining evidence centered design (ECD) and the types of justifications and documentation that are required for documenting validity of inferences at multiple steps along the test development and analytic cycle, Mislevy and his colleagues describe an open system that conceptually can handle different types of evidence for different sets of test takers (e.g., Mislevy, Steinberg, & Almond, 1999; Mislevy, Steinberg, Breyer, Almond, & Johnson, 1999, 2001). They focus on explaining the points at which evidence is required, rather than the types of evidence needed for different types of students. Even with this limitation, their work provides a solid foundation for building frameworks that would justify when and how to consider variations in testing conditions.
- Different students interact with items and test conditions in different ways depending on the active characteristics in items and the ancillary abilities of students. This point becomes the primary lens through which decisions need to be made about how task demands are presented to students, how students interact with tasks demands, and what conditions might facilitate this exchange.

Recent advances in cognition provide a basis for thinking how students approach, address, integrate and retrieve concepts and skills. Further, research supports that students move through identified sequences in different ways and at different rates, depending on a multitude of attendant individual and environmental factors. However, much of the work on task processing and learning seems to have been focused on qualities of the tasks and student competence regarding the desired task targets, rather than on identifying unique and common characteristics in students and how these interact with processing in a variety of tasks. For instance, Embretson (2003) identifies task-specific cognitive sub processes in tasks, and models how students respond to the different sub processes so that their varying scores are a function of their degree of mastery in the target abilities. Lohman and Bosma (2002)

point out that both experimental/cognitive and differential/measurement psychologists frequently array their data in a person by task matrix, and that both groups of psychologists have tended to emphasize the main effects in this matrix. While experimental/cognitive psychologists emphasize differences among tasks/treatments and differential/measurement psychologists emphasize differences among persons, both desire to minimize the interaction between persons and tasks.

There is some work that attempts to explore the person/task interactive space. In a major review, Pellegrino, Baxter, and Glaser (1999) focused on intelligence and aptitude tasks, explaining how the cognitive components approach worked to develop differential models of task performance by exploring components of performance that varied across individual task takers. The approach assessed performance strategies, executive routines, and examined how targeted declarative and procedural knowledge interacted with the varying processing capabilities of task takers.

Some researchers explore the targeted and construct irrelevant aspects of the task/construct space while generalizing across students. Snow and Lohman (1993) appear to draw a distinction between component skills and strategy adoption, suggesting the separability of perception, memory, verbal and special abilities (as well as strategy) from the targeted construct. Glaser and Baxter (2002), among others, define a content-process domain space for school science within which targeted constructs can be classified and defined in terms of task types. The four quadrants of the space (content—rich to lean, process—constrained to open) provide a context for differentiating targeted construct aspects of task performance from construct irrelevant aspects, if such item characteristic differentiations are specified.

As Kopriva, Winter and Wiley (2004) discuss, for the purposes of creating test tasks and testing conditions that work for a variety of students, including those English language learners with low proficiency in English, the focus is the encounter of individual test takers with test tasks and testing conditions presented and implemented in specific ways. The question is under what conditions target skills are properly conveyed and when communication about targeted information becomes systematically contaminated, misunderstood, or distorted. The proper function of the ancillary components is to act as a facilitative agent. However, they may also act as a barrier, and this barrier can be considered to be an issue of person/task *access*. The nature of the interaction between the active components in items and conditions and the ancillary abilities of individual students needs to be evaluated before decisions can be made about how the task demands should be presented to the students, and acted on by the students. Specifically, issues of access need to be considered at three points: a) understanding the meaning of the task demand, b) allowing the individual's problem solving mechanisms to operate within the task demand/condition space (e.g. identifying problem solving strategies, selecting and assembling the proper strategies, and implementing the strategies to come to a solution), and c) allowing students to retrieve, organize, and demonstrate their task solution.

- Kopriva (1999) presents a framework for the comparable inclusion of students with diverse needs and challenges into mainstream large-scale testing programs. The paper lays out the philosophical and methodological basis for comparable inclusion in such a way that, given proper validation arguments and empirical evidence, one could justifiably aggregate scores over students who take the test under different conditions.

Essentially, she suggests that advances in evidentiary reasoning (e.g., Schum, 1994) and statistical modeling (e.g., Gelman et al., 1995) allow us to bring probability-based reasoning to bear more flexibly on the problems of modeling and uncertainty that arise naturally in all assessments. Developments in instructional and cognitive psychology make it clear that students access and process different information, and utilize personal internal schemas for organizing and demonstrating their responses regarding this information (e.g., Greeno, Collins, & Resnick, 1997). Researchers in these fields (e.g., Glaser, Lesgold, & Lajoie, 1987) have proposed models for how to harness these individual differences in group classroom and large-scale assessment settings. Further, more precise item modeling approaches have focused on understanding the deep conceptual structures within valued academic content constructs, and ancillary components that affect the measurement of knowledge and skills in items and tasks (Haertal and Wiley, 1993; Wiley & Haertal, 1994).

Kopriva argues that the work mentioned in the previous paragraphs makes it possible to extend principles that underlie familiar test theory in ways that may be able to provide more accurate and nuanced inferences from sets of data. That is, advances make it possible to explicate and test alternative grounds for inferences from assessment data. The standard argument for common inferences has been made on procedural grounds: common content in items and a common approach for synthesizing and summarizing items and response data over items. The latter part of this argument required standardized conditions of observation as a key aspect of synthesizing item data. However, based on developments in fields such as those mentioned above, we can now develop, implement, and test an alternative conceptual argument for common inferences. As in the procedural argument, the measurement of common substantive content is important. But rather than requiring standardized conditions of observation, the conceptual argument can be built on evidencing appropriate inter-relationships between target inferences, the knowledge and skills of interest, necessary observations, the properties of tasks or items designed to elicit the observations, and the assessment situations where students interact with assessment requests. This approach suggests that data may be collected under alternate conditions.

At the core of any framework which considers the use of condition variations is a clear and precise understanding of task targets at the item level. Further, the core must elucidate not only the active properties of items designed to elicit the intended observations, but also active properties of items which are irrelevant to the intended targets. All items include both sets of stimuli for all test takers. While the intent is to present the same test targets to all students, the conditions in items which explicate these target task demands contain stimuli which interact differently with different

students. Ideally the target-irrelevant or ancillary properties function as a facilitative agent, however, sometimes they act as a barrier to accessing the target demands for some test takers. Kopriva (2008) explains the nature and characteristics of ancillary item properties.

Kopriva (1996) distinguishes between central ancillary components and non-central ancillary components, where the central components are those which are not part of the construct, but are usually prerequisites to targeted knowledge or are used in the demonstration of that knowledge. For example, to solve an algebra problem in mathematics usually requires using arithmetic as well as algebraic skills. Central components may also be certain concepts associated with language, procedures or algorithms. Non-central ancillary components include elements such as knowledge or communication and contextual skills that are needed to answer the question, but are not materially related to the construct. They also include cultural interpretations associated with presenting stimuli, including words, contexts, or procedures. Kopriva suggests that non-central ancillary components are always irrelevant to the target. On the other hand, central ancillary abilities are sometimes considered to be relevant (they would not be distinguished from the target demands in the observed score), and sometimes irrelevant to the target. It is important for item writers to make decisions about the nature of an ancillary component because, if and when it is considered to be target irrelevant, access variation supports can be legitimately used. Even when the ancillary components are considered to be target relevant, it is useful to identify their presence, since they become essential building blocks for operationally defining and restricting the item target construct.

The following equation represents the impact of the various active components in items on the observed test scores of students. y represents the item score, c represents the influence of the target component, or what is intended to be measured, while a_k and b_l represent the central and non-central ancillary components, respectively. i and j represent items and students, respectively. e is the random error component.

$$(1) \quad y_{ij} = c + a_{1ij} + a_{2ij} + \dots + a_{kiu} + b_{1iu} + b_{2ij} + \dots + b_{liu} + ab_{1iu} + \dots + bc_{1iu} + \dots + ac_{1ij} + \dots + abc_{ij} + \dots + e$$

$$= c + \sum_i a_{kiu} + \sum_i b_{lij} + \sum_i ab_{iu} + \sum_i bc_{ij} + \sum_i ac_{ij} + \sum abc_{ij} + e$$

Note that the influence of a_k and the b_l can represent either main effects or interactions with the item target and/or with other ancillary components. In the latter case, the effects of c and a_k or b_l on y are interdependent. Furthermore, it is recognized that the target-irrelevant components can function both conjunctively and in a compensatory fashion. Thus, the score variance is influenced by target and non-target elements that covary within and across target, and by ancillary elements.

Equations 2 and 3 illustrate one way in which the ancillary and target variables may interrelate conjunctively and in a compensatory manner, respectively

(2) (placeholder)

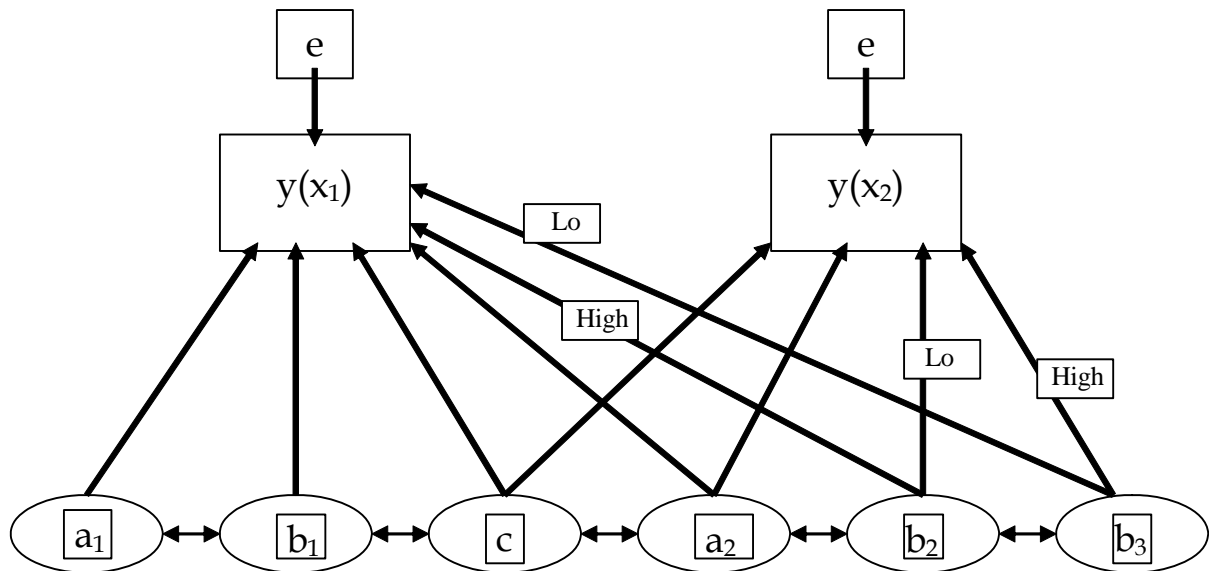
(3) (placeholder)

The goal is to have students interact with items and test conditions in such a way that the active characteristics associated with the item target (c) are maximally reflected in y and that any active characteristics that are irrelevant to the target are present in a facilitative function, rather than acting as a barrier to the student interacting with the targeted task demand.

Actually, of course, the relationships of abilities or skills and knowledge sets to observed scores is compounded by measurement errors of the causal latent factors as well as by random and unidentified systematic errors associated with observed item scores. The c , a_{kij} , and b_{lij} terms impact the observed score (y) for each student. Further, the observed score is always a function of the conditions under which it is measured. Therefore, y_{ij} is actually $y_{ij}(x_{mj})$. $y_{ij}(x_{mj})$ 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_m . The set of conditions (presentation, administration, and response) are those test taking situations under which the task demands are introduced to the student. 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.

The observed score, depicted in this fashion, is illustrated in Figure 1. Here, y is observed under two conditions, x_{1j} and x_{2j} . The same c is conceptually being measured by both conditions. Three b 's in addition to the a_{kij} , and c , impact $y(x_{mj})$. The non-central ancillary variable b_{l2j} , which corresponds to the same measurement condition x_{2j} , impacts $y(x_{2k})$ highly, where the opposing b_{l2j} does not. b_{l1j} and a_{1j} influences x_{1j} , and, thus, because of the measurement conditions, the two measurements of y are impacted differently as well. Other interactions are possible, even probable, but are not presented here. As noted above, the relationships of one variable to a second one will often have an effect on other variables and so on, and this is also not developed in Figure 1.

Figure 1



The discussion of active components in items/tests can also proceed from a linguistic perspective, in particular a linguistic perspective sensitive to the semantic and functional values attaching to linguistic elements. Meaning and function, of course, also attach to non-linguistic knowledge representations of the sort prominently featured in dynamic test items such as those in the ONPAR project. Since functional linguistic perspectives themselves are motivated in cognitive and communicative realities, such a perspective could in principle generalize at some level to include both linguistic and non-linguistic representations. In any event, the perspective needed to invoke must be rich enough conceptually and terminologically to effectively characterize the active item components as core or ancillary and as essential or non-essential. Note that these characterizations of active components presuppose a notion of functional distribution, so the choice of a functional linguistic framework should theoretically be able to incorporate the distinctions.

A Linguistic Model to Investigate Item Components

The linguistic perspective invoked here in inform the following discussion represents an amalgam of sorts of two theoretical and descriptive traditions in linguistics, both of which extend to incorporate language behaviors at a discourse level. The first and older tradition is referred to *functional grammar* and the other as *cognitive linguistics*. Both traditions have been advanced in different ways and with different points of view by various researchers. Rather than pick and choose from among these different perspectives, though, two of the most fully elaborated and cohesive theories in each tradition were selected. In the functional grammar tradition, the target theory is the one advanced over more than fifty years of work by M.A.K. Halliday and his co-workers and most

comprehensively described in Halliday and Matthiessen (2004). In the cognitive linguistics tradition, the target theory is that advanced by Leonard Talmy under the more specific name *cognitive semantics* and most comprehensively described in Talmy (2003). Cognitive semantics has itself a long development history and may be looked on as synthesizing work in a number of earlier and still independent traditions in linguistically oriented semantics, lexical studies, pragmatics, and cognitive psychology.

It is beyond the scope of this paper to detail either of these traditions, but both have provided essential insight and guidance in structuring the following arguments. In particular, certain distinctions and terminology of the operational framework, shown diagrammatically in Figure, have been borrowed directly from one or both traditions. For instance, the characterization of discourse structure as comprising *experiential*¹, *interpersonal*, and *textual* functions comes directly from functional grammar, as do the characterizations of the general situational context of a discourse in terms of *field*, *tenor & mode*, and *mood*. Likewise, the characterization of the experiential function of discourse as comprising *schema configurational*, *perceptual*, and *attentional* components comes from the cognitive semantics tradition as do the terms *windowing* and *perspective* and the notion of *warrants* and *salience*. Other specific terms carrying semantic weight, such as *register*, *participant relations*, and *cohesion* are widely embedded in many semantic traditions and have even entered the general language of education and assessment in ways that are congruent with their specialized semantic senses.

These notions have been compiled and organized in Figure 2 to lay a basis for investigating the claim that the representations in dynamic, computer-based items, such as have been developed for ONPAR, make similar content demands on the test taker as do the representations in traditional, static items. A further purpose is to provide a model for better identifying and describing the differences between the ancillary components of items produced in these contrasting models. The comparison model rests on three assumptions: first, that the parallel, ONPAR versus traditional test items both constitute examples of human discourse; second, that both discourses are predicated on common semantic cores stemming from a common cognitive framework; and third, that the realizations of the parallel items – one largely language based, the other largely graphic – functionally share essential organizational and meaning components.

The model structures a claim that a test item entails a communication between a test giver (or item writer) and a test taker. This communication forms the basis for the characterization of the item as a discourse and motivates the inclusion of the *interpersonal* function in the model. There are expectations on the part of both these participants in the discourse that will ideally abide by the guidelines governing all well-formed and well-executed discourse. These guidelines are most succinctly expressed in

¹ Halliday and Matthiessen (2004) use the term *ideational* to name the metafunction, here labeled as experiential. *Ideational* in their usage incorporates experiential and logical discourse functions (p. 29). The breakdown is intended to highlight a difference between the objects of experience (experiential) and the relations that obtain between these objects (logical). However, for purposes of the arguments advanced here the distinction is overly nuanced and the term *experiential* seems more accessible and apt, given the instructional context in which testing is embedded.

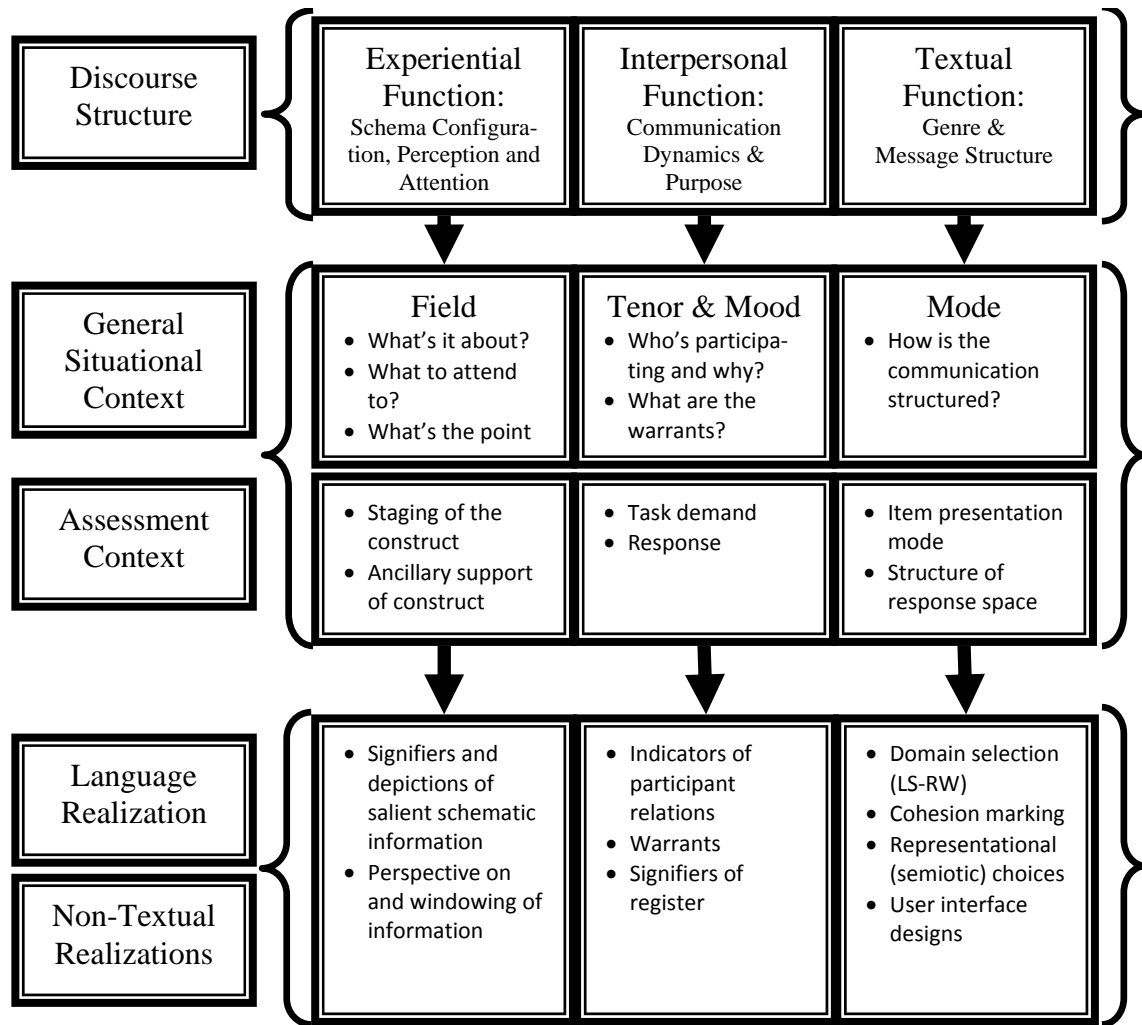
Grices's (1975) *cooperative principle*, which states that participants in a conversation or text exchange will cooperate with one another in assuring that the message of the communication is transmitted accurately, succinctly, relevantly, and appropriately for its purpose. These four desiderata for a successful communication are encoded in what are called *conversational implicatures* expressed by maxims of quality, quantity, relation, and manner, respectively.

As a test item is concerned, the conversational implicatures pertain to the purpose of the item, that being to determine the test taker's proficiency in a well-delineated area of content. The item writer communicates a request for information from the test taker for an appropriate demonstration of that proficiency. This is to be accomplished within the constraints set up by the item writer's choice of item type and by some statement pertaining to the scope of the required response. These requirements are controlled through meanings conveyed via the textual function of the discourse (the third column in Figure 2).

The intents of the item writer and of the test taker are expected to align. In a typical conversational exchange the intents are made clear or are implied through warrants of the truth value of the communication and of the right of the participants to participate in the communication. The test taker must accept the warrant that the test giver proffers that he or she is entitled to request information, while the test giver must accept the warrant that the test taker is providing a good faith response. The details of these warrants are more often tied to claims of status and authority existing outside of a particular exchange, but the language of the exchange, through the adoption of particular communication registers, will often buttress and support these claims. Meanings of this sort are conveyed through the interpersonal function of the discourse (the second column in Figure 2).

Finally, the test item itself is aimed at a well-conceived construct that becomes operationalized in the item—together with the request for specific, related information—as the assessment target, or item target, at a smaller grain size. The assessment target focuses and sets bounds on the construct, which lies within the scope of the test taker to engage; that is, it lies within his or her experience. Essentially, then, the test item predicates a schema and poses a demand for the test taker to demonstrate some skill or knowledge relevant to that schema. The meanings attendant to this part of the discourse are described in the experiential function (the first column of Figure 2).

Figure 2 Linguistic Framework for Describing Active Elements of Test Items



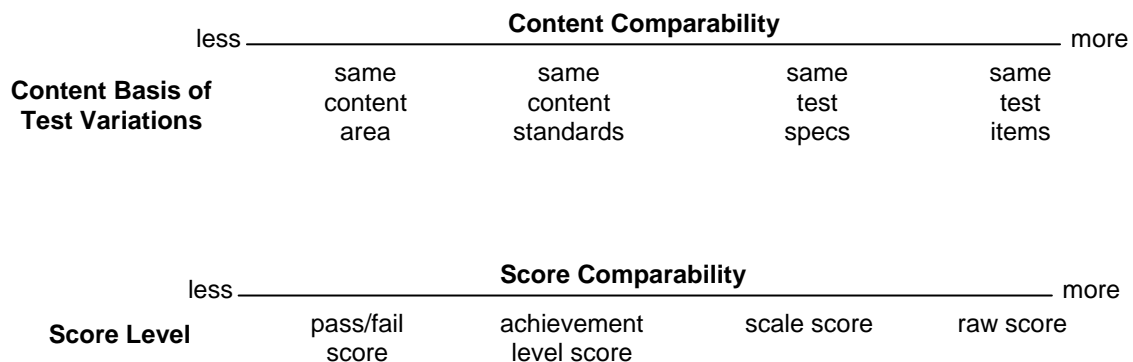
Violations of the cooperative principle can take place either through a willful action on the part of one or both participants, or more typically in a testing context, by some unintentional means. For instance, an intentional violation of the relation maxim would occur if the item writer wrote an item he or she expected no test taker could successfully respond to. While it's reasonable to claim that malicious intent is rare, it does happen, and quite often, that test items are given to students who lack the ability to comprehend the task demand or to respond, not by virtue of their lack of knowledge of the construct, but because of the ancillary demands of the item—those that concern in this example the choice of an inappropriate language or register in which to communicate the item demands and in which to produce a requisite response. The consequences of such violations are well known, of course, and they have in particular motivated the current effort to produce a science test that is more equitable for ELLs. In any attempt, though, to ameliorate a poor communication channel—as in this example, by reducing the language load—a risk is incurred that other elements necessarily introduced to compensate for the reduced language will themselves become causes of bias. Hopefully, by being more

cognizant of the communication implications of various item components, we will ultimately be able to create items that we can be more sure will adhere to the cooperative principle.

The ONPAR Projects

In order to study the possibility of moving large-scale testing from a static to a dynamic interactive testing environment two ONPAR projects have been funded to date. These projects are focusing on improving the measurement of academic knowledge and skills of students with low English proficiency by investigating selected questions. In particular, they are researching what elements need to be included in academic items built to be used in a dynamic environment, so that they can be used to suitably measure academic content of students in a large-scale situation. Some item differences, for instance, language of translation (when applicable) will change over students whereas other item elements will remain the same over students. Because of the nature of the interactive environment and the nature of the students' language and cultural challenges, multiple item types beyond those used in traditional testing will be used. Further, the projects are investigating how well the same targeted skills can be measured in these items as compared to their static counterparts, and when and to what extent the two sets of items can be considered comparable. Figure 3 outlines the content basis of test variations, and the score level of comparability that might be considered if proper evidence is obtained. While the projects are interested in keeping the same targeted content and cognitive complexity levels across the static and dynamic items, there will be differences in item types between static and dynamic items measuring similar content, and differences in cognitive loads of content-irrelevant stimuli. Therefore, project investigations are aimed at collecting evidence related to the scale score comparability level. The science ONPAR is using science items from the NECAP testing program, and NECAP states and testing staff are providing guidance for this project. The mathematics ONPAR will be focused on mathematics items, and a number of static item sources are being considered at this time.

Figure 3 Adapted from Winter, 2007



The research agenda for the two grants will use cognitive labs during item development to understand how ELLs and native English speakers are interacting with the dynamic items. Independent judgment reviews of both the static and dynamic items will provide

evidence that similar item targets are being measured across item pairs, or not, and they will identify cognitive demands in each set of items that are non-essential to the science targets. Two sets of controlled trials using NECAP and ONPAR items will be completed during the science ONPAR project. These trials will randomly assign NECAP or ONPAR items to native English speakers and ELLs of different proficiency levels. ONPAR forms will examine how item elements common to two versions will impact the measurement of targeted stimuli for the different groups of students, and how students will interact with different levels of language and translations in the two versions. Finally, it is expected that a large scale study will be conducted with the static and ONPAR mathematics items which will investigate the convergent and discriminant validity of the items.

Primary Components of ONPAR Items

A major focus of the ONPAR projects is concerned with thinking through what elements must be included when well-conceived assessment items are developed and presented in a dynamic format. This is a large undertaking as it entails a detailed task analysis of the standard base items and detailed task design development work for each new ONPAR item. The task analyses of each of the standard items involve the ‘deconstruction’ of the active characteristics in order to identify the item target and cognitive complexity inherent in the academic demands, other relevant constraints associated with the item, and other active elements irrelevant to the targeted demands. The task designs for the ONPAR items involve several iterations. First, the concept is articulated, including how the base item target and targeted cognitive complexity is to be re-constructed for ONPAR. The initial task also involves identification of target-irrelevant elements that the staff believes would aid in the facilitation of the task demand for the population of low English proficient ELLs. Following the conceptual explanation, the ONPAR draft item is then storyboarded and refined through reviews by project staff. Next, the ONPAR item is programmed, reviewed, and revised. Until staff are clear about what elements allow students to focus on the task demands, items are tried out in the cognitive labs and revised as appropriate. Following is a brief summary of some of the primary elements associated with ONPAR items.

First, the ONPAR items enlist the dynamic computer capabilities of simulation and animation to present primary task contexts, including depictions of movement and time-lapse, and also to present secondary contexts relevant to the task demands. These primary and secondary contexts are formatted to clearly present salient information, omit distractive elements, and place related supportive visual cues in the background. In static formats these functions would have been realized primarily through text. Recently some items on traditional tests have used static visuals to mirror text or to replace portions of it (see Kopriva, 2008, Chapter 6), and, sometimes static visuals are included here as well. For the most part, such depictions interact receptively with the test taker.

Second, the ability to manipulate stimuli that have been systematically presented on-screen allows students to interact expressively with the task. This includes interacting with stimuli during problem solving, and demonstrating responses through manipulation of non-text or minimal textual elements. Interacting with computer capabilities allows

students to pursue such activities as point and click selection, drag and drop, modeling, assembling and dis-assembling, connecting and dis-connecting, using freehand drawing, and engaging slider bars and roll-overs, all for the purpose of answering the targeted questions. In an effort to understand the students' conceptual thinking, some items ask students to explain phenomena by manipulating conditioned sets of text, visuals, and symbols where the sets are underpinned by if/then, and and/or logic.

Third, ONPAR items may also use the computer capabilities of visual, audio, and in some cases, bilingual roll-overs as support for language that remains. These types of supportive elements are becoming increasingly common in online testing, which takes static items from a paper and pencil test and places them on the computer. Besides roll-overs, text supports also include use of arrows and other directional information that orient and focus the student's perspective. Whereas traditional online large-scale tests may use some of the aforementioned computer capabilities to systematically produce accommodations for students, and minimize the numbers of interfaces the student interacts with during testing, ONPAR goes beyond these functions. The ONPAR approach introduces supports that interact inextricably with the other digitized components of ONPAR items, supporting the identification of task demands, contexts, and interactive elements, and becoming part of the arsenal of interactive stimuli used by the student.

Fourth, just as paper and pencil static tests rely on certain fonts, page formats, and other standardized presentation and administration elements that occur in test booklets across items, ONPAR is in the process of identifying, developing and using over-item standardized elements that are most salient for this dynamic testing situation. The function of the paper and pencil elements seems to be to provide a predictable, undistracting framework within which the student can focus on the item requirements. Likewise analogous elements would serve the same functions in ONPAR. To date, such elements include a standard set of animated icons representing different process demands; standardizing frame size of the item, placement of item stimuli, and response spaces; and standardized navigation and item "maintenance" operations such as, clearing response spaces, submitting responses, and activating animations.

Fifth, a decision was made to constrain the scope of the items in one particular way. That is, when a standard item asks the student to choose the correct characterization of a scientific or mathematical word or phrase, ONPAR writers are dividing this task into two portions, a conceptual portion and a definitional portion. ONPAR items will handle the conceptual aspect of such items. However, the item writers will leave the definitional portion to be separately implemented by educational agencies outside of the ONPAR core of items. In that way, the agencies could add a static item measuring whether a student can define the academic word. In this way item targets focused specifically on technical vocabulary can be addressed in a way that will highlight the difference between the student's conceptual grasp of the science versus his or her proficiency in the English based definitions of technical language.

Item Types Used in ONPAR

Research has suggested that traditional multiple choice items have several drawbacks for ELLs, especially those with the lowest level of proficiency in English. Using a ‘shorthand’ textual approach in stems and particularly options choices is often problematic for this population who rely on context to understand the language. Further, for several cultural and experiential reasons, the technique of responding by choosing an answer among a set of pre-established options is frequently not the best method for understanding what these students know. Rather, direct demonstration of their solution seems to be preferred as does providing them the opportunity to explain themselves. While the object of the ONPAR project is to provide a close-ended environment for responding to items, one intent of the work is to investigate item types that may be effective for this population while still being computer scorable and comparable to the kinds of items used by the mainstream population. The set of item types identified so far as candidates for ONPAR items is detailed in Table 3. These item types use the features noted above and those summarized below. They are meant to work with the array of skills and cognitive complexity measured in large scale tests, including those measured by multiple choice and constructed response items and ranging from the measurement of simple recall skills to demonstrations of students’ higher order conceptual processing. It is anticipated that the item types could also be used with more extended tasks.

Table 3 ONPAR Identified Item Types

<p>Traditional Item Types</p> <ul style="list-style-type: none">• Multiple Choice• Free/Open Response <p>Complex Computer Dynamic Item Types</p> <ul style="list-style-type: none">• Select and Classify• Matching and Connecting• Ordering Stimuli• Stimulus Manipulation: Simulation of context and target stimuli to represent item requirements; students manipulate stimuli to address requirements• Statement Construction (Simulated Constructed Response): Students construct syntactically ordered statements using visual and linguistic elements; statements may include Boolean or conditional logic• Representational Modeling: Modeling using toolkit elements; incorporates assembling and drawing• Graphical Modeling: Modeling of quantitative relations by manipulating a graphic representation <p>Composite Item Types (multi-tasked items with tasks of different item types)</p> <ul style="list-style-type: none">• Thematic Probing: Loosely integrated tasks each leading to a discrete result• Problem Solving Vignette: Tightly integrated, temporally sequenced tasks converging to a single result
--

Questions about Amount of Language and Use of Translations

In this section two approaches are considered for representing test items, both with the common goal of reducing the language load of test items. The two approaches, which are here designated *Low Language* and *Very Low Language*, represent two characterizable stages on a continuum of language load. It has already been mentioned that many traditional paper and pencil items are also incorporating other non-text based representations. It's not clear whether such developments are specifically aimed at the goal of reducing language load—though a case could be made that it does do so. In any case, such traditional and near-traditional items lie on the high side of the language load scale. The goal of the ONPAR project is to effectively test how low on the scale one can go while still providing sufficient supports in alternative ways.

The same demands face the ONPAR item writer as face any item writer operating in a traditional paper and pencil environment; that is, to make the task requirement or item question unequivocally apparent to the test taker, to provide a suitable environment and mechanism to allow students to solve the problem, and to provide avenues for the test taker to demonstrate his/her solution. For the target group of ONPAR test takers, it is generally felt to be insufficient simply to state or depict the demand because of their severe English limitations and translation problems; the demand must minimize the need for language and support the remaining language with ancillary devices. What is not clear, though, is the extent to which those ancillary supports must be explicit or may be inferable. The two positions considered each make different arguments about the need for explicitness.

Both of the options are characterized by some common features explained above in the primary components of ONPAR items.

Using Low Language and Translations

The first approach argues that the ultimate goal of a test item is to make the item demands and avenues for problem solving and responding explicit and very clear to the students. This approach maintains that the item target and related directional steps should be spelled out precisely so that a student does not have to rely on inference to understand what the item is asking him or her to do. Further, it is argued that an amount of redundancy, using both translation and non-textual elements, is considered important to support the language, as long as it is not too verbose, distracting or overwhelming.

- Because of the diversity of the ELL population, expectations relative to all aspects of the items and tests need to be clear. An understanding of expectations cannot be taken for granted as cultural and regional differences can vary a great deal for different reasons. Item writers may think making explicit the type of information discussed here is too obvious and a waste of time and resources.

Farr and Trumbull (1997) and Hiebert and Calfee (1990) discuss the need to be direct, clear, and specific about what the item requirements are, what latitude the student has in solving the problem or responding, and how students can use additional time, tools, or other accommodations. Clarification of non-targeted vocabulary is essential, but

should not substantially increase the language load. Malcolm (1991) argues that contextual parameters associated with pre-requisite knowledge expectations and response constraints should be explicit and clear. For instance, if the student is supposed to anticipate certain consequences or ways of viewing phenomena, these should be explained in a way that does not dramatically increase the literacy load.

- The target item requirements need to be clearly cued using language as well as non-textual elements, where, in turn, the text and non-textual elements could sometimes mirror and support the meaning of the other. The specificity of task demands needs to be clearly stated using as morphologically and syntactically simple language as possible while still retaining precision and specificity of meaning.
- Irrelevant information which enhances the task demands and forms the context for the item requirements should act as a facilitator to the target content. While non-textual animation and simulation, as well as the iconic elements, will often act out or suggest their roles within an item, they should be supported by enough language so that a student does not have to expend inferential resources on trying to figure out why these elements are included in the items. It is expected that there will be some redundancy in how text and non-textual elements interrelate.
- Because of the novelty of the item types and the dynamic nature by which the items are presented to the students, it is important to be very explicit about how the student is to move through the item. That is, the mechanics and procedural implications of the individual item approaches should be very transparent to students. This includes information relative to understanding how the target requirements unfold in an item, relevant information related to accessing any and all tools relevant for solving the problem, and information which articulates the avenues of response open to the students using the particular item approach. Until students are used to interacting with these types of items, it may be argued that the item writers should err on the conservative side and support each step with enough language as it takes to clearly convey the mechanics of the approach. It is understood that the language, for the most part, will be mirrored by other non-language visual and manipulation elements, by support tools such as rollovers and audio, by animation such as that which portrays time passing and interactions, and by activities that require students to interact with data elements prior to the primary test question. Redundancy in item elements associated with the procedural aspects is most likely seen as beneficial, at least initially, as students become very familiar with these item approaches.
- It is acknowledged that the use of translations are not ideal as they do not often exactly mirror the English, and for some languages and in some instances the translation may even be problematic for particular words or concepts in individual items. In large part, ONPAR items were created in order to address this disadvantage through greatly reducing the amount of language to begin with and substituting computer capabilities such as animation and interactive stimuli for much of the language found in standard items. However, once these substitutions are made, it is argued that the remaining language should be supported with translations, except in

cases where the non-text elements can clearly, and unambiguously, stand on their own. It is understood that this increases the redundancy, but until the tradeoffs are understood, this should remain a preferred method.

Using Very Low Language in English

Following are a few considerations associated with the argument of using as little language as possible in ONPAR and restricting this language to English.

- Keep language constructions simple. Language demands present themselves at all levels of functional linguistic organization: experiential, interpersonal, and modal, and all organizational levels are supported by lexical, morphological, and syntactic choices. The very low language option tries to reduce the influence of the morphological and syntactic levels by presenting item demands using individual words or word clusters, rather than relying on phrase and sentence formation rules of English syntax. That is to say that units of language more complex than the individual word are not used. In addition, morphologically complex words are avoided.
- Avoid all L1 translation. The anti-translation case argues that translations may adversely affect or affect in unpredictable ways the staging of an item's construct. Languages do not contain the same or even parallel devices to represent the configurational structure of an item's content. Consequently, it could happen that a translation introduces an unanticipated variation in how the construct is represented. To the extent that the visual and animation components of an ONPAR item reflect English discourse structures, it could happen that a translated text suggests other discourse structures in such a way that they could cue an answer. Also the practical goal of trying to maintain standardization across twenty or so languages is daunting.
- Make the task demand inferentially clear. The greatest challenge of the very low language approach is to make the task demand apparent to the test taker using almost no language. When the language load is very low, the burden of making the demand unequivocally clear falls on the power of other structures in the item to communicate that demand. In the very low language version of the test items, presuppositions (i.e., knowledge of word definitions and associations) essential to full understanding of the item requirements are realized through roll-overs establishing word-to-visual or visual-to-visual object links.
- Keep information redundancy low. Recent studies in the field of cognitive loading (e.g., Wallen, Plass, and Brünken, 2005), have shown that providing multiple sources of written annotation to science text increases cognitive load on comprehension and can reduce comprehension. For the ONPAR target group, phrasal and sentence based English itself constitutes a redundant element, since targeted test takers are unlikely to control the lexicon and grammar needed to use the text as a primary source for understanding. Including such text will be arguably distracting, rather than facilitating, to item comprehension.

References

- Embretson, S. E. (2003). *The second century of ability testing: Some predictions and speculations*, Princeton, NJ: Educational Testing Service.
- Farr, B. P. & Trumbull, E. (1997). *Assessment alternatives for diverse classrooms*. Norwood, MA: Christopher-Gordon Publishers.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. London: Chapman & Hall.
- Glaser, R. & Baxter, G. P. (2002). *Cognition and construct validity: Evidence for the nature of cognitive performance in assessment situations. A festschrift in honor of Sam Messick*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Glaser, R., Lesgold, A. & Lajoie, S. (1987). *Toward a Cognitive Theory for the Measurement of Achievement*. Pittsburgh, PA: Learning Research and Development Center, University of Pittsburgh.
- Greeno, J.G., Collins, A.M., & Resnick, L.B. (1997). Cognition and learning. In D. Berliner & R. Calfee (Eds.), *Handbook of educational psychology* (pp. 15-47). New York: Simon & Schuster Macmillan.
- Grice, H. P. (1975). Logic and conversation. In Cole, P. and Morgan, J. L. (Eds). *Syntax and semantics: Speech acts*. Volume 3. (pp. 41–58). New York: Academic.
- Haertel, E. H., & Wiley, D. E. (1993). Representations of ability structures: Implications for testing. In N. Frederiksen, R. J. Mislavy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 359-384). Hillsdale, NJ: Erlbaum
- Halliday, M. A. K., & Matthiessen, C. M. I. M. (2004). *An introduction to functional grammar* (3rd ed.). London: Hodder Arnold.
- Hiebert, E.H., & Calfee, R.C. (1990). Classroom assessment of reading. In R. Barr, M. Kamil, P. Mosenthal, and P.D. Pearson (eds.) *Handbook of research on reading* (2nd edition) (pp. 281-309). New York: Longman.
- Kopriva, R.J. (1996). *Variant Methodology for Different Testing Populations*. Paper presented at a meeting for Meta-SCASS, Washington, D.C.
- Kopriva, R.J. (1999). A conceptual framework for for the valid and comparable measurement of all students. Unpublished paper for the Council of Chief State School Officers.
- Kopriva, R.J. & Cameron, C. (2007) Comparing standard and enhanced access items for diverse students: Item analyses in six grades and four subjects. Presentation at the

CCSSO Large Scale Assessment Validity and Evaluation.

- Kopriva, R. J. (2008). *Improving testing for English language learners*. New York: Routledge.
- Kopriva, R.J., Winter, P.C., & Wiley, D.E. (2004, April). *Rethinking the role of individual differences in educational assessment*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.
- Lohman, D.F. & Bosma, A. (2002). Using cognitive measurement models in the assessment of cognitive styles. In H. Braun, D. Wiley, and D. Jackson (eds.), *Under construction: The role of constructs in psychological and education measurement* (pp.127-146). Hillsdale, NJ: Lawrence Hall Associates.
- Malcolm, S.M. (1991). Equity and excellence through authentic science assessment. In G. Kulm and S. Malcolm (eds.), *Science assessment in the service of reform* (pp. 313-330). Washington, DC: American Association for the Advancement of Science.
- Mislevy, R.J., Steinberg, L.S., and Almond, R.G. (1999). On the structure of educational assessments. *Measurements: Interdisciplinary Research and Perspectives*. 1(1), 3-62.
- Mislevy, R.J. Steinberg, L.S., Breyer, F. J., Almond, R. G., and Johnson, L. (2001). Making Sense of Data from Complex Assessments. CSE Technical Report
- Pellegrino, J. Baxter, G. P., and Glaser, R. (1999). Addressing the “two disciplines” problem: Linking theories of cognition and learning with assessment and instructional practice. In A. Iran-Nejad (ed.) *Review of research in education* (pp. 307-353). Washington, DC: AERA.
- Schum, D. (1994). *The evidential foundations of probabilistic reasoning*. New York, NY: John Wiley & Sons.
- Snow, R.E. and Lohman, D.F. (1993). Cognitive psychology, new test design, and new test theory: An introduction. In N. Fredericksen, R.J. Mislevy, and I.I. Bejar (eds.), *Test theory for a new generation of tests*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Talmy, L. (2000). *Toward a Cognitive Semantics*. Cambridge, MA: MIT Press.
- Tindal, G. (2006). The journey through the reliability of a decision-making model for testing students with disabilities. Presentation at the annual meeting of the American Educational Research Association, San Francisco, CA.
- Wallen, E., Plass, J.L., & Brünken, R. (2005). The Function of Annotations in the

Comprehension of Scientific Texts – Cognitive Load Effects and the Impact of Verbal Ability. *Educational Technology Research and Development*. Special Issue: Research on Cognitive Load Theory and Its Design Implications for E-Learning, 53(3), 59–72.

Wiley, D.E. & Haertal, E. (1995). Extended assessment tasks: Purposes, definition, scoring and accuracy. In R. Mitchell (ed.), *Implementing performance assessment: Promises, problems, and challenges*. Hillsdale, NJ: Lawrence Erlbaum Associates.