Chapter 12: Where Are We and Where Could We Go Next?

Summary and Next Steps

Rebecca Kopriva
University of Wisconsin

This handbook has provided a useful template for thinking about measurement comparability and it has identified several useful methods for potentially demonstrating the extent of comparability between test variations and general large-scale tests. The present chapter will summarize selected points, placing them within a context of history and focus, what can be done today, and future possibilities and challenges.

As Winter mentioned in her introduction, over the last 15 to 20 years, rethinking how to best measure academic content led to rethinking what aspects of test-making, implementation, and analysis need to be involved in order to make large-scale test form variations comparable at some level of interchangeability. Specifically, in the 1990's advances in cognitive learning theory led to the identification of an expanded set of measurement approaches that seemed to be promising for use in large-scale assessment. The advances included re-considering what were valued constructs associated with academic content, and identifying novel approaches associated with how students might demonstrate their capabilities in large-scale academic settings. Such approaches included a) the use of open-ended writing prompts designed to measure the writing skills of students; b) short answer and more extended open-ended items meant to measure the students' sophistication in understanding content, such as mathematics, or history, through explaining their conceptualizations, reasoning or meta-cognitive skills; c) performance tasks

Kopriva, R.J. (2010) In P.C. Winter (Ed.), *Evaluating the comparability of scores from achievement test variations*.

where students manipulated materials and demonstrated their skills first-hand; and d) development and use of student work portfolios meant to chronicle academic maturation using multiple methods. As these approaches were tried out, evidence was accumulated about what types of knowledge and skills were measured with various formats, how these methods might be developed to withstand use across classrooms, schools, and districts, how they might be administered in a variety of school contexts, and how results should be analyzed. Various challenges, some sizable, were also noted. Concurrently, measurement experts began to think about how to determine when and how different kinds of performances could be seen as 'comparable'. The focus in those days was on comparability of responses *within* approaches, for instance, when rubrics allowed for various ways for students to demonstrate their knowledge and skills at, say, a level of 2 or 3 out of 4 possible points. The focus was also on determining the kinds of standardized procedures that needed to be in place to build viable performance tasks or constructed response items that measured different content in different academic domains; determining how to construct defensible rubrics for various types of writing prompts, constructed- response items, performance tasks, or portfolio evaluation systems; and identifying standardized criteria that were needed to properly constrain the student work entries in portfolios while still allowing for variability in the types of entries. Because of competing costs and other issues, work on the approaches has largely stalled for the last 10 years or so. However, the efforts left behind a legacy of work with direct implications for expanding how comparability might be considered, and what types of evidence need to be in place to make this claim.

Current Comparability Challenges

Recently, the clearest challenge to using one set of methods and form types for all students comes from broader inclusion requirements associated with today's statewide academic assessment systems. The 1994 reauthorization of the Elementary and Secondary Education Act required more inclusion of students with disabilities (SDs) and English learners (ELs) in their accountability reports. In 2001 the No Child Left Behind (NCLB) followed the lead of the earlier reauthorization and the wide-ranging students with disabilities legislation (IDEA) passed in 1997. It mandated specific accountability requirements aimed at not only inclusion but explicitly improving the academic performance of these two populations. NCLB also expanded the reviews of the statewide assessment systems to more closely evaluate whether or not the systems were defensibly measuring the academic achievement of SDs and ELs. The focus on special populations highlighted that, for some of these students, traditional large-scale testing methods were not sufficient or valid in assessing their knowledge and skills (i.e. see Kopriva, 2008). Not surprisingly, this focus resulted in an increase in the development of other methods that were supposed to be more appropriate for these students. The methods included adaptations of administration conditions involving the use of accommodations, construction of different kinds of forms, and alternative kinds of procedures aimed at collecting academic information in ways that vary from the general testing formats used for most students. Some questions of comparability involved determining the viability and interchangeability of performances within the testing system which were obtained under varying administration methods. Other considerations, including the work contained in this handbook, focused mostly on comparability among different form variations.

Within the last few years more attention has been paid to assessing in order to support learning as opposed to just what students have learned, doing so in a way that is responsive to identifying errors, while also being defensible, consistent over classrooms, and often over schools (for instance, see Pelligrino, Chudowsky & Glaser, 2001). This charge highlights the variability in learning processes over students and over time. As the field moves forward, viable learning-sensitive assessments will need to be receptive to different progressions as well and wrestle with questions of comparability within students at different points and across all students, not just English learners or students with disabilities.

Carlos Martinez, former associate superintendent in New Mexico, recently posed a pointed question: Do or should testing inferences change when the makeup of item types within a test radically changes from one year to the next? To afford the cost of large-scale academic assessments or to otherwise address changing standards or other pressures, in the last few years several states have altered the manner in how they assess their school populations. Most often, the systems have moved from more to less or no constructed-response items in their testing systems or in particular content areas. Some work has been done to consider how item type interacts with targeted constructs and therefore with inferences that can be made about student mastery (e.g. Li, Ruiz-Primo, & Shavelson, 2006). Over a series of papers, these authors and others identified the four salient links between item aspects and valued knowledge or skills. They suggested that the first two item aspects, task demands and cognitive demands, established a link that differentiated question type by valued learning while the additional two, item openness and additional factors, served as mediators to the link. Considering comparability over time when testing systems change their item types, or when significant weighting changes occur within tests is to-date something that has not been researched. However, system alterations such as those suggested here appear to have comparability implications as well.

Issues of comparability are not 'going away'. Rather, as a field, it seems that increasingly we are being held accountable for what we say are the test score inferences, as they pertain to students with particular profiles and, soon, students with different kinds of learning experiences. This handbook provides some guidance about how to think about comparability when variations exist in the system. Systemic frameworks such as Mislevy et al.'s (e.g. 2003) Evidence Centered Design could provide some direction about what elements in the development, implementation and analysis of assessments need to be present, which need to be standardized, and what evidence at each of these steps needs to look like or include. Comparability, like validity, will be an argument to be made and criteria of which arguments are suitable for which purposes will need to be identified and tested.

The Comparability Questions

To begin the conversation, Winter essentially identified three questions that need to be addressed as we consider comparability:

1. What do we want when we want score comparability?
2. What do we mean when we say comparability?
3. How can we evaluate comparability?

The first question seems to focus on the inferential achievement claims the test evidence can support. Evidence will come from elements in the development, implementation and analysis of performance data. Documentation of the procedures used to produce the evidence will need to pass scrutiny and should be evaluated through the lenses of appropriateness for capturing the knowledge and skills of particular students in particular situations—In other words the evidence is viable if the procedures address and minimize alternative explanations. It is probable that test score evidence will come, to a reasonably large degree, from viable evidence at the item level, or other kinds of criteria or constraints meant to focus the types of responses required about particular content and skills.

Winter's grain size discussion seems to address the second question. That is, is comparability focused at the scale score, achievement score, or single cut-point level? This makes a difference for the kinds of evidence that need to be collected, with the overall expectation that scores from both the general test and variation should be considered 'interchangeable enough' and without flags. If the focus is one cut-off score (as in pass-fail), the whole assessment exercise should be focused on producing performances correctly identified on one side of the cut off or the other. If more than one but a discrete number of scores are of interest, then interchangeability documentation needs to address the same question at each of the cut points. When raw or scale scores are the focus then evidence needs to demonstrate that multiple scores along a continuous range are measuring similar enough knowledge or skills for the students taking each form.

The third question focuses on how to analyze the evidence and make decisions about if the documentation is ‘good enough’. As Winter points out, there must be sufficient evidence of both construct equivalence and score equivalence. Basically, construct equivalence focuses on grounding the meaning of the score inferences resulting from students taking the test variation or the general test, and making sure the user can have confidence that the meanings are the same (or the same ‘enough’). This aspect of equivalence reflects the analysis of evidence produced to defend question 1. Score equivalence focuses on documenting that the scores from the variation and the general forms are behaving in the same way (or the same ‘enough’) for students with similar abilities. Evidence that will be analyzed for this aspect of equivalence comes from data which are appropriate to address question 2, that is, to defend the claims of interchangeability at the targeted grain size.

Examples of construct equivalent evidence that need to be evaluated include:

- same standards coverage
- similar criteria for inclusion
- similar judgments about relevant cognitive demands
- similar internal structure

Evidence of score equivalent evidence includes:

- similar proficiency percentages
- similar score distributions
- similar rank order

Properly used, evidence of score equivalence at the item level may be demonstrated by methods such as

- similar distractor distributions
- similar DIF results
- similar p-values.

It is important, particularly at the item level, that score procedures and group samples be suitably scrutinized and vetted to minimize the possibility of alternate inferential explanations for one or more groups under study.

On the whole it seems that, as variations are more divergent from the general test in terms of format and approach to collecting information, evidence of construct equivalence between the variation and the general test need to be more stringent at a particular grain size. Further, the notion of scores used for what purpose should probably also be considered here—higher stakes at whichever grain size suggests more defensible evidence that alternative explanations for the performances have been considered and found to be not tenable.

As a field we need to get better at linking inferential claims to the methods we are using to collect the data about students' knowledge and skills. Currently research is lacking in when and how methods changes affect the targeted constructs 'enough', and if and when non-targeted cognitive demands are a factor and for whom. As the changes over years question posed by Martinez is debated, and as more variations in large-scale testing occur, it will be imperative that these distinctions become clear. Then we can confidently claim that one set of differences reflect somewhat distinct, but similar, inferences where the same may not be true for another set. Finally, Dr. Winter has suggested that the criterion for making interchangeable inferences is to focus on the sparsest method used in the testing system that is part of the comparability argument. In other words, portfolios, if properly assembled and evaluated, would probably yield a great deal more information about student skills on targeted constructs than would a multiple choice test. However interchangeable inferences would be tied to the multiple choice form, perhaps with supplementary information available from the analysis of the portfolio contents.

Efforts We Can Make Today

The work on broadening what we mean by comparability started nearly two decades ago. From Messick's work (see Messick, 1989) and the research and development work in the 1990's we reminded ourselves of the importance and centrality of validity in determining the rigor of tests and their interpretations, and lately we have come back to the notion that comparability is essentially linked to validity. We rediscovered the item as a basic unit of analysis in validity, the importance of being clear and explicit about intended item-level targeted cognitive knowledge and skills, and the need for evidence to back up the intentions (not just vague notions at the subtest or test level). We also connected evidence arguments to validity and by extension comparability. Technological progress allows us to increase the standardization of all conditions and has shown promise in enhancing how some students can acquire meaning of items and be able to tell us what they know. These advances may be useful in bridging how students with different profiles can differentially take and respond to tests in ways that engender confidence in common inferences over different methods. All in all, several steps have been taken which can be directly applied today to improving how we consider and evaluate comparability among form and administration variations. Among them five are highlighted here.

1. Operationalizing the Evidence Centered Design Principles and Applying Them to Test Variations

Mislevy and others' Evidence Centered Design or ECD (e.g. 1996, 2003) presents a detailed framework for identifying the units and functions associated with developing assessments and specifying how test components must be synchronized to effectively support the interpretations of scores. Besides identifying particular tasks and procedures within a test and its parallel forms, ECD can also be used to build a comprehensive assessment system structure that concurrently plans for and considers design and evidence elements of all variations as well as the general test. One of the most central and important aspects of ECD is that the inferential claims expected from the test and subtests should be identified first, not last. Once these warrants are clarified, then the design of the system, the test frameworks, procedures and development of items are undertaken to be commensurate with producing data to satisfy these claims.

Another advantage of using ECD is to outline a string of evidence that demonstrates when a traditional system is robust. For instance, among other types of standardized elements, Evidence Centered Design suggests the use of task templates from which to build more traditional item versions. In these task templates contexts, specific elements associated with the targeted construct, characteristics of option choices, certain parameters, or other aspects are purposefully changed in order to ‘grow’ the task pool. These items may be used to build other forms in the assessment system or used for other purposes. ECD can also be used to design a system where testing situations change because of student challenges, or it might be useful, potentially, when more than one learning map within a progression necessitates changes in tasks or relationships between tasks. Conceptually, ECD can handle systems where very different methods of collecting student performance data are included as well.

An important part of ECD, especially when test variations are part of the system, is the explanation of alternative arguments for certain students, and clearly specified options plainly linked to students with specific needs. Take for example the situation where variations are being proposed to improve accessibility. First, before options are identified, test developers need to consider arguments about why the general test (including the presentation of forms and general standardized administration and response conditions associated with it) isn’t ‘good enough’ for particular groups of students. These arguments should propose alternative explanations about why students might be scoring as they do, explanations that are due to other factors than the intended constructs. As such, they call into question the meaningfulness and accuracy of the proposed general test score inferences for these students. For the alternative explanations to be effective they need to be tied to students with particular profiles, for example students with the lowest levels of English proficiency and little first language literacy. The precision at this step is essential so that testing solutions can be clearly identified which propose to minimize the alternative argument and bolster support for the overall inferences of the assessment system. A more thorough explanation of how alternative arguments need to be defined can be found in Chapter 12 of Kopriva (2008).

Second, procedural materials need to be proposed and evaluated to determine if they are viable and feasible candidates for minimizing the alternative arguments *and* supporting the general score inferences. As an example, solution options might include a signed version of the forms for deafand hard-of-hearing students, or a translated script and an oral first language administration of forms for low English proficient students with little first language literacy. Third, evidence needs to be collected that the content constructs have not been altered by the options as compared to what was intended on the general test. Fourth, oversight documentation needs to be collected that the options were implemented in a standardized fashion—so, for instance, were steps taken to avoid cuing in the administration options by developing signing videos so students across the state would hear or see the same administration? Fifth, additional evidence should be collected reflecting that the variations performed as intended, including data that the options and the general test reflect specified similarities necessary if the administrations or forms are to be considered comparable. This type of evidence would probably include checks that both the general and option conditions produce similar structures of responses, including similar factor loadings or other types of dimensional scaling data.

Thus, for any type of variation, it is important to identify the target group who will benefit, identify the points within item, task and/or test development, administration, scoring or analysis that seem to be causing problems for the targeted group, and, within the points, what in particular is problematic. Once this is established proposed solutions that are salient for the targeted group need to be identified, and implemented, and appropriate follow-up needs to occur to provide evidence that the variation is behaving in a 'similar enough' fashion relative to the general population on the general test. As a general rule of thumb, whenever anything is changed at any point within development or administration, there needs to be adequate evidence that the solution is working as expected for the targeted group.

All in all, Evidence Centered Design can provide construct equivalent documentation, and this type of 'conscious' design will also probably have an influence on producing 'good enough' evidence of score equivalence. It can be used to identify where evidence needs to be collected, and what types of evidence might be viable and possibly necessary in making a strong argument that the inferences are comparable for everyone taking a test within the assessment system.

2. Defining Suitable Elements

In most cases when form variations are proposed there are changes to the language approach and/or the method by which data are collected. Even when relatively straightforward 'clarified language', translated, Braille, or ASL signed forms (completed visually) are suggested, each of those approaches has its own challenges. Testing methods variations, for instance portfolios or performance events, including scripts and forms to complete that document how students demonstrated pre-specified skills or knowledge, present sizable changes to how questions or tasks are presented to students. Each of these sets of changes necessitates that suitable guidelines be specified to guide the development of the variations to produce forms of similar scope. The changes also necessitate that the developers design and implement a plan to independently evaluate and document if the completed variations and general forms are measuring ‘similar enough’ topics at similar levels of cognitive complexity. Additionally, identifying and evaluating form components such as these extend to the selection and use of supplementary tools. These might include, for example, bilingual glossaries, highlighters, or procedural guidelines associated with administration and response accommodations. They also extend to a robust and suitable explanation of how students with identified challenges are matched to accommodations they need, and to oversight procedures to ensure that the matching implementation went according to plan. As the federal peer review guidance has begun to evaluate, these types of evidence as well as evidence documenting fair use of accommodations assigned to those taking the general test vs. those who receive no accommodations are all part of building and demonstrating adequate construct equivalence documentation.

When methods differ substantially from general testing forms and procedures, they almost always involve additional cognitive demands. To date, there has been little work isolating which cognitive demands are salient and should be common across methods. The Barton and Winter chapter specifies several points during development and implementation that they suggest need to be evaluated, and last section below will briefly outline some of the challenges inherent in how clarifications and guidance about these topics might be specified.

Besides the identification of suitable components and procedures, the specification of suitable relationships between elements/procedures also seems to be essential. This extends to decisions of score equivalence (which scores or other quantifiable relational indexes need to be 'similar enough' across options), and also to decisions of what is considered suitable data to begin with. That is, for variations where data collection components are different than general test components (e.g. items) or vary substantively by type (e.g. type of items—multiple choice vs. constructed response), when can components be compensatory and for whom, and when do conjunctive rules apply? When is it ok for students to demonstrate a skill one time? When should the instrument require more evidence? Initially these are judgment calls. Assuming common inferences across students taking different types of forms, methods, or types is the goal, when structured analyses of relationships of different salient variables and their magnitudes are completed, they will help inform the parameters we can comfortably work within. Today, it is suggested that we begin to frame the questions, isolate the components, elements, and relations of interest, and produce some rudimentary evaluations of our decisions.

3. Identifying the Grain Size of Targeted Components

In order to determine construct equivalence, one important aspect is to specify the data points in the variation and general test that will be compared. While the final focus of equivalence is usually at the test level (across what are generally recognized to be parallel forms), is it good enough to only address construct and score equivalence at the aggregate score test level? When item-by-item versions of general test items are made in a variation, several researchers suggest an evaluation by general-variation item pairs to determine if they are measuring the same intended content targets (for instance see De Pascale in this volume, and Sireci and Wells' reference to producing adequate translations). These independent judgments would be a necessary but not sufficient addition to post hoc analytic confirmation of similar test structures and other evaluations of score equivalence at both item and test levels (e.g. see Sireci and Wells in this volume).

But what about when the nature of the item or tasks is different over the general test and variation purported to measure the same content and cognitive skills? For portfolio entries Rigney and Pettit (see below) suggest clearly and specifically defined criteria. To evaluate the similarity of targeted content and cognitive complexity, should the set of entries corresponding to each of these criteria be the grain size of the portfolio variation that should be compared to the corresponding set of items on the general test? Who should perform this evaluation—is it 'good enough' for internal reviewers to document this crosswalk or should an independent judgment by a credible panel or external organization make the judgment calls? Since the general test and variation are supposedly measuring the same content standards, it would seem reasonable that both sets of data collection schemes would be evaluated against the intended constructs in the content standards. The result is essentially another type of alignment document.

The recent work on producing the alternate assessments for states' most cognitively disabled students suggest techniques for producing defensible evaluations of testing methods, although, they are pegged to different content and achievement standards. This means that the only comparison with the general test is at the achievement standards level. How might the comparative methods used in this context be extended to variations when the same content standards are also in play? The reason this question is important is that the alternates are expanding the kinds of defensible data points other variations might use, including, for instance, the use of observational protocols, performance events and other methods designed to directly collect information about students' knowledge and skills. How are these approaches constrained in development and implementation to produce responses the field is comfortable with, and what kinds of evidence needs to be documented to support the inferential judgments?

Initial answers to the questions posed above span a range of research that is currently available (for instance see the literature reviews associated with the studies cited in this volume). Many challenges remain as more development efforts share their findings, but some basic approaches are available now and can be more widely utilized.

4. Standardizing Development and Implementation Elements

In his chapter, DePascale gives examples of the type of development procedures testing systems might employ when clarified language form variations are being used as part of the assessment system. The Barton and Winter chapter summarizes several aspects of development, implementation and oversight which can and should be standardized and evaluated when test variations are substantially different than the general test. They also provide examples of the kinds of evidence that might be collected when these elements are in place. For instance, it is apparent that the target constructs needed to be well-defined, clear, and narrow in focus, and that clear 'alignment' techniques need to be put in place to determine if the general test and the test variations are measuring similar information. While the content standards provide the basis for this work, the burden of proof is on the variation and therefore more specific guidelines about how to evaluate data from this approach relative to the general test would seem to be warranted. Gong and Marion (2006) discussed the tradeoffs of building assessments where the items, form elements, and even accommodations are more standardized up front (as in traditional on-demand tests) versus the trade-offs when flexibility in entries is different across students. When different types of data collection methods are part of the overall system, these kinds of tradeoffs need to be considered when the comparability arguments are presented.

To provide credibility of the test variations themselves several elements need to be in place. Rigney and Pettit (1995) reported that identification of clearly defined characteristics of entries in portfolios was essential, which do not specify the nature of the tasks but require that specific representation elements need to be included in order for scorers to properly evaluate the work. For some approaches where teachers and students supply work, completion of a few anchor tasks may be required for all students. These common anchor tasks help to calibrate or anchor the scores across students. Additionally, strong, clear scoring rubrics, and in-depth rubric notes were key to building a successful variation, with training of scorers consistent with the standards of the field. Producing a defensible approach that could be systematically scored across students, teachers, and schools requires forethought and planning. Auditors of the statewide language arts portfolio system in Kentucky found that criteria could be effectively communicated throughout the state, and that teachers could learn to be accurate scorers, given proper training and rigorous oversight. The state of Maryland found that rigorous auditing with samplers and training of proper techniques was successful at minimizing variation and drift when locally evaluated elements were part of the testing system (Ferrara, 1999).

In Rigney and Pettit’s report portfolios were used for all students, rather than the use of portfolios as a variation of the general test and designed to be used for only some students. However, the document suggests the type of rigor that would need to be in place for the variation to confidently yield defensible data. Then, in addition, the use of portfolios as a variation would seem to require further criteria so evidence could be built to document if the variation was performing ‘similarly enough’ to the general test.

5. Using Simultaneous Item/Task Design

About 15 years ago Texas undertook an interesting approach to item development. Items in English were developed, reviewed etc. and then items in Spanish were translated from the English. While still in the pre-operational stage, if translators/ Spanish item writers found that they could not suitably translate the English versions, and changes suitable to both versions could not be found, the English items were discarded. In other words, items had to adequately convey the same meaning in both English and Spanish for this pair to be part of the state’s large-scale test. This is a rudimentary example of simultaneous item development.

Using Evidence Centered Design, as variations become a stable part of testing systems (including the use of accommodations as well as form variations or other alternative types of data collection methods), approaching development and implementation in a systemic way involving all parts of the system seems reasonable. Rather than a “do the general test first and then ‘jerry-rig’ variations to address special populations” approach, simultaneous design considers the entire system from the beginning, builds various kinds of items/tasks measuring the same targets, and assembles a thoughtful set of procedures designed to convey the targeted meaning of the items/tasks to students with particular profiles and receive responses from them that are meaningful. A priori designs of analyses implemented in order to confirm the give and take of meaningful content knowledge and skills to and from the developers and students produce evidence of the level to which the methods are working as intended. For today's large-scale systems it has long been argued by some (for instance, Kopriva, 1999, 2008; Tindal & Fuchs, 1999) that actually there should probably not be one test (the 'real' test) with accommodations (not as good as the 'real test' but what can you do…). Instead, a better approach would be different but equal methods to measuring 'similar enough' content and skills, each with some sense of which students would benefit from each particular method without causing undue advantage. Further, as we gear up for the multi-dimensional benchmark tests and other systems designed to focus on supporting ongoing learning in a reliable way over classrooms and for student profiles with different learning maps, it would seem that simultaneous test design approach would be well suited. In either case, ECD provides the framework. As large-scale test implementation migrates to computers, algorithms designed to handle and direct various conditions to students who need them would support this type of development. To date it seems that the main drawback is lack of a cohesive vision and probably reluctance to spend the time upgrading well-worn development procedures to consider the full spectrum of work up front.

Looking Ahead

1. Unpacking Comparable Demands for Tests that Support Learning

As noted above, there is growing interest in developing interim or benchmark assessments. In balanced assessment systems these assessments sit at a middle ground between informal and formal classroom formative tasks occurring in an ongoing fashion throughout the school year, and summative tests focusing on what students have learned over usually a year-long period. Mark Wilson (2004, 2008) suggests that there are two types of possible benchmark tests. One mimics a summative assessment and is focused on an evaluation of education, albeit over a quarter or semester as opposed to a year. The other type is designed to evaluate the end-status of education but also to provide guidance for continuing instruction based on demonstrated errors in conceptualization or level of skill development. It appears that most consumers would prefer the second type of assessment over the first, but limited understanding about how to build them has hindered their development on a large scale basis. In particular, clear and definitive learning maps keyed to construct grain sizes that are useful for this kind of work are in short supply, as developers struggle to avoid the minute progressions associated with learning research, and instead establish a manageable number of key indicators that demarcate levels of knowledge and skills maturing over time.

While levels along the continuum from novice to expert are difficult enough to identify, it is even more of a challenge to reasonably understand the pathways related to how students with different profiles might move through the levels. As the various pathways are conceptualized, tasks which differentiate levels and also various student progressions are being built. These tasks, and the assessments which embody them, will need to be scrutinized to determine if they are providing reasonable access to students from diverse profiles, and if common inferences across pathways at the assessment level can be supported by adequate evidence. In this case, the profiles, pathways and the access they require won't necessarily be by student subgroup (for instance, an EL or a student with a disability), but will be dependent, to a reasonable degree, on previous learning opportunities interacting with other schooling and personal experiences.

For these kinds of assessments, at least three types of comparability questions appear to be important. First is a key concepts question: "What are the salient levels of the learning maps for a particular scope of study and what are the key pathways between and across levels that adequately capture learning progressions for the full range of students?" Second is a substantive question: "How should different pathways be reflected in different versions of tasks and sets of tasks as necessary, concurrently retaining the integrity of the common constructs being measured across versions while possibly collecting different kinds of information or collecting information in different ways?" Third are two related evidence questions: "What are suitable types of evidence for documenting the veracity of the progressions and pathways, and for documenting access for the full range of students? What kinds of score equivalencies, and at what grain sizes, are acceptable for which purposes?"

It seems that the comparability work completed today will inform these complex questions tomorrow. Currently researchers are finding that access issues can be conceptualized with a discrete and parsimonious number of profiles, and that comparability concerns often can be addressed on the basis of these discrete profiles (for instance see Carr & Kopriva, 2009). While the profiles for the supportive assessments will be different, most likely, there will also be a reasonable and parsimonious number of them that can adequately capture the broad scope of student learning within a scope of topics and cognitive demands. Further, work on understanding under what conditions different methods can yield comparable information will inform this work as well. As a field, learning from the current comparability challenges, we will be much closer to understanding how to address the comparability needs associated with these new types of assessments.

2. Taking Advantage of Interactive Computer Capabilities

Every day we are becoming more technologically savvy, and expanding our use of computers for presenting and scoring a broader range of assessment tasks will be no exception. Already several states and large-scale testing systems have moved their traditional testing systems online, including Minnesota who has incorporated contextual sequences of animation into their discrete large-scale science items. In addition, there are notable examples of fuller assessment tasks in science, including Quellmalz and others' work (2007; 2008) with embedded interactive classroom tasks, and NAEPs experience with computer-based interactive extended tasks. Kopriva and others (2008; 2009) have demonstrated that using multi-semiotic interactive representations to replace language in items for low English proficient students are successful in producing scores on par with those of native English speakers taking either the same interactive items or their traditional item equivalents.

Using the capacity of computers expands how item contexts can more fully engage students in assessment as compared to how traditional tests typically present their questions. This capacity extends how tasks can build on and intertwine evolving contexts with test questions by using animated or simulated sequences, and video and/or audio clips. Computer-based assessment can also access the internet or selected archives of data or related sources, systematically broadening the scope of what information students might be able to use to solve problems or answer questions. Additionally, of course, using the computer and its various capabilities increases the kinds and complexity of algorithms we can easily use to score items for summative use, or elements of items within and across tasks for more formative purposes. All in all the use of computers as a delivery and response capturing tool opens up multiple avenues for enhancing large-scale testing, avenues formerly available only for classroom or research purposes.

Nonetheless, as we are pondering the potential for expanding how we might use computers to assess student knowledge and skills, it is also clear there are differences in how students negotiate these types of interactive items compared to their traditional static counterparts. Issues of comparability would seem to be important to consider in reconciling when data from these types of tasks are compared to or interrelated with data from traditional formats. Of these, four seem to be immediately relevant.

First, many interactive items use visual displays of contexts to provide the backdrop for target questions. Although these contexts are computer-simulated, often they are reminiscent of the 90's performance tasks where students would be presented with concrete materials and equipment with which to conduct an experiment in science or collect data to be analyzed in mathematics. As compared to traditional items which explain contexts using words or static drawings or photographs, these interactive items appear to engage the students more directly in the intended construct and the relevant surround, seemingly decreasing the distance between the latent construct targets associated with the tasks. This difference between the latent targets and the manifest ways they are being measured could have comparability implications. Kopriva, Gabel, and Bauman (2009) have referred to the distinction between static and interactive items as a difference in “stickiness”, where it would seem that the greater cognitive engagement in the interactive items could have either a positive or facilitative effect, if communicated well, or a negative, perhaps confusing, effect when delivered poorly. To-date understanding when greater and lesser distance between latent and manifest targets affects measurement and how has been largely unexplored.

Extending the response opportunities in assessment tasks is another capability eminently feasible as computers are used to deliver tests. To-date, other than extending the ability of students to complete tables, almost all of the projects still require that students respond to interactive tasks using traditional close-ended responses or typed-in constructed-response explanations. However, Kopriva and others have experimented with expanding how to capture performance responses directly, allowing students to manipulate stimuli within the tasks in order to demonstrate their reasoning and skills. For Kopriva et al., these advances are essential because their target groups of students with substantial language challenges find traditional methods of response to be language intensive and impossible to navigate, especially as more complex content is assessed.

However, the implications for the work would clearly appear to extend to all students, because capturing performance responses would seem to broaden the range of latent cognitive schemas that could be accessed—schemas that the students' executive functions may not yet have had time to meta-cognitively process into language. Just as the impact of interactive contexts have not been determined or mapped, the same is largely true here. For what types of questions might performance responses be most useful or effective? If students are allowed to respond differently, when do differences matter and when don't they? How should these questions be answered when testing purposes are different, or when students have limited access to traditional methods of response and must use other avenues?

Third, these types of simulated and interactive tasks seem to involve greater amounts and kinds of cognitive demands than do static items. Cognitive psychologists point to the increase in the *density* of the cognitive demands, which appear to reflect increased connections of external and internal stimuli. As compared to static traditional items, in some cases density in interactive items appears to be related to minimizing the distance between the test questions and the latent constructs being measured; sometimes it may be associated with other elements within or across students' internal cognitive maps. The density phenomenon also seems to reflect quantity of input which, depending on students, the effectiveness of presentation, and other conditions, could either lead to a fuller, more complete picture of the task (think about more pixels per square centimeter), or to stimulus overload. To some extent the measurement field must not only become more aware of the distinctions between construct relevant and irrelevant elements of items, but also of how these distinctions are made manifest through greater or lesser density.

While elements of static items often either were or weren't facilitative for various populations, density in these 'three dimensional' tasks probably involves continuums where even the most facilitative aspects might reach a threshold beyond which the demands are too imposing. As this is understood, how density interacts with both the targeted content and a broader range of student experiences and abilities will need to be considered as well. Advertising and other mixed-media fields have learned how to use density to their advantage; likewise, as test developers, we need to learn how and when density (or lack thereof) is useful or problematic in producing the targeted and facilitative effects we intend.

Fourth, how the relevant cognitive schemas are engaged in animated and interactive tasks appears to differ in some cases relative to how the schemas are engaged in static representations. To date most research on the relationships between schemas and test questions has focused on using language (in text or orally) as the primary conveyer of meaning, although sometimes language might be supplemented with other elements such as static drawings or visual representations (Kopriva, Winter, & Wiley, 2004; Kress G. & van Leeuwen, T., 2006). However, research related to how students with language challenges (such as those with disabilities, English learners, and probably poor readers) learn as well as cognitive science advances (e.g. see National Research Council, 2001) makes clear that there are alternative avenues of comprehension and acquiring meaning. Based on this, it seems reasonable to assume that interactive assessment tasks probably do sometimes stimulate some other internal maps associated with meaning than the types of items we currently use. For many students, meaning could be activated through either language or other methods so the net effect may be the same.

For those with substantial language difficulties, alternative methods activate essential meanings in the assessment task so students can access the item question where otherwise they couldn't. Further, if effective response possibilities are present, interactive elements in the items allow them to respond with what they know as well; if they aren't available the lack could negate the good intentions associated with other access improvements. What we don't really know is when or how we might be triggering different or even possibly conflicting methods of conveying meaning for different students and if or when this might be problematic. As one example, Carr (2006) found that certain elements in some types of static visuals were differentially 'read' by English learners in one way, by students with learning disabilities in another way, and by deaf and hard-of-hearing students in yet a third. Clearly, this implication needs further study.

Comparability questions in today's assessments seem daunting enough; challenges inherent in the considerations associated with future work appear overwhelming. While it may be tempting to revert back to "the good ol' days", a caution is that much of the work we are facing today and tomorrow is actually work that *should* have been done years ago. That is, although we blithely assigned inferences to test scores for all test takers, we understand now that we perhaps weren't actually as clear about what we were actually measuring as we purported to be. Yes, we had learned the importance of standardizing conditions, checking for reasonable consistency over time and within tests, and understanding how we might design items and evaluate the responses to generalize over students with various abilities. Yes, these attributes greatly improved the robustness of the score interpretations we made. However, we also knew of the high correlations between social economic status and total scores (and while there was certainly evidence to support that students with more resources learn more and better, they were probably too high); that our item development processes weren’t always as cognizant as they should have been about the inferences we were making; and that our item type restrictions were sometimes limiting what we could actually measure. So, as we move forward in considering how to properly measure concepts and skills for the broadest range of students, it is time to catch up with the research in the learning sciences and consider how to more thoughtfully build items and tests to differentiate novice and mature learners. Overlaying technology on these findings could and should, then, allow us to reflect back to teachers more information about how to guide instruction, and at the same time use the computer capabilities to align more defensibly with the richness of constructs underlying content in the various subject areas. The test construction and awareness of test taker diversity that we address today will provide the necessary foundation for tomorrow’s advances.

References

Carr, T.G. (2006). Application of STELLA system and relevant findings. Paper presented at the annual National Conference on Large-Scale Assessment, San Francisco, CA., June.

Carr, T.G. and Kopriva, R.J. (2009). It's about time: Matching English learners and the ways they take tests by using an online tool to properly address individual needs. Paper presented at the National Council of Measurement in Education, San Diego, CA., April.

Ferrara, S. (May 28, 1999, personal communication).

Gong, B. and Marion, S. (2006) *Dealing with flexibility in assessments for students with significant cognitive disabilities (Synthesis Report No. 60)*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Kopriva, R.J. (1999) Making State Tests inclusive for special populations: Training guidelines for developing and implementing Inclusive Title 1 Assessments. Washington, DC: Council of Chief State School Officers.

Kopriva, R.J., Winter, P.C., and Wiley, D.E. (2004) 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, April.

Kopriva, R.J. (2008). *Improving Testing for English Language Learners: A Comprehensive Approach to Designing, Building, Implementing, and Interpreting Better Academic Assessments,* Routledge Publishers, NY, NY.

Kopriva,R.J., Gabel, D. and Bauman, J. (2009). What happens when large-scale items actually use computer capabilities? Exploring issues and redefining challenges. Paper presented at the National Council of Measurement in Education, San Diego, CA., April.

Kress, G. and van Leeuwen, T. (2006). *Reading Images: The Grammar of Visual Design*. Routledge Publishers, London, England.

Li, M., Ruiz-Primo, M.A. and Shavelson, R.J. (2006). Towards a science achievement framework: The case of TIMSS 1999. In S.J. Howie and T. Plomp (Eds.), *Contexts of Learning Mathematics and Science: Lessons learned from TIMSS*, Routledge Publishers, NY, NY.

Messick, S. (1989) Validity. In R.L. Linn (ed.), *Educational measurement* (3rd edn) (pp. 13–103).Washington, DC: American Council on Education and National Council on Measurement in Education.

Mislevy, R.J. (1996). Test theory reconceived. *Journal of Educational Measurement*, 33, pp. 379-416

Mislevy, R.J., Steinberg, L. and Almond, R. (2003). On the structure of educational assessments, *Measurement: Interdisciplinary Research and Perspectives,* 2003*, 1,* 3-67

Pellegrino, J.W., Chudowsky, N., and Glaser, R. (eds.) (2001) *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Quallmalz, E. (2007). Calipers: Using Simulations to Assess Complex Science Learning, abstract received from http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=0814776

Quellmalz, E., DeBoer, G., and Timms, M. (2008). Foundations of 21st Century Science Assessments, abstract retrieved from http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=0814776

Rigney, S. and Pettit, M. (1995) Criteria for producing equivalent scores on portfolio assessments: Vermont's approach Presented at the annual meeting for the American Educational Research Association, San Francisco, CA, April.

Tindal, G. and Fuchs, L.S. (1999) *A summary of research on test changes: An empirical basis for defining accommodations.* Lexington, KY: University of Kentucky, Mid-South Regional Center.

Wilson, M. (Ed., 2004). *Towards Coherence Between Classroom Assessment and Accountability.* National Society for the Study of Education Press, Chicago, IL.

Wilson, M. (2008). The nature of quality benchmark assessments. Keynote presentation at the Center for the Assessment and Evaluation of Student Learning Conference, San Francisco, CA, October.