Re-Examining the Role of Individual Differences in Educational Assessment

Rebecca Kopriva
David Wiley
Phoebe Winter

University of Maryland College Park

Paper presented at the Annual Conference of the National Council of Measurement in Education, April 2004, San Diego, CA. Work based in part on research funded by the U.S. Department of Education, Grant No. R305T010846-03 and Grant No. 5368A030005.

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 sets 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, Pelligrino, 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 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.

The general view of the conceptual approach described here is that the central aspect of any testing situation is a person interacting with a test task. This interaction seems to be a neglected area of study. We believe that, because of the inactivity, sometimes significant amounts of systematic error have been misidentified and have led to distortions in test scores and in score inferences of some students. Our goal is the understanding and modeling of this interaction process space, with special attention paid to how we can minimize the effects of irrelevant factors when they interfere with the measurement of targeted knowledge and skills. We suggest that this interference, when it occurs, is actually a function of aspects of the person/task interaction, when irrelevant characteristics in items interact with sub-par skills in some students.

Additionally, it seems that work such as Snow and Lohman (1993) and Glaser and Baxter (2002) could open the door for developing more individually tailored items and assessment approaches keyed to measuring the same target abilities while minimizing the effects of the irrelevant factors. In this case, we are interested in understanding how tests designed and developed to maximize the measurement of targeted knowledge in the person/task interaction space can defensibly yield common inferences and comparable scores across persons being assessed in varying ways. This work involves understanding the types of evidence that will be needed to make sustainable arguments about comparability and common inferences.

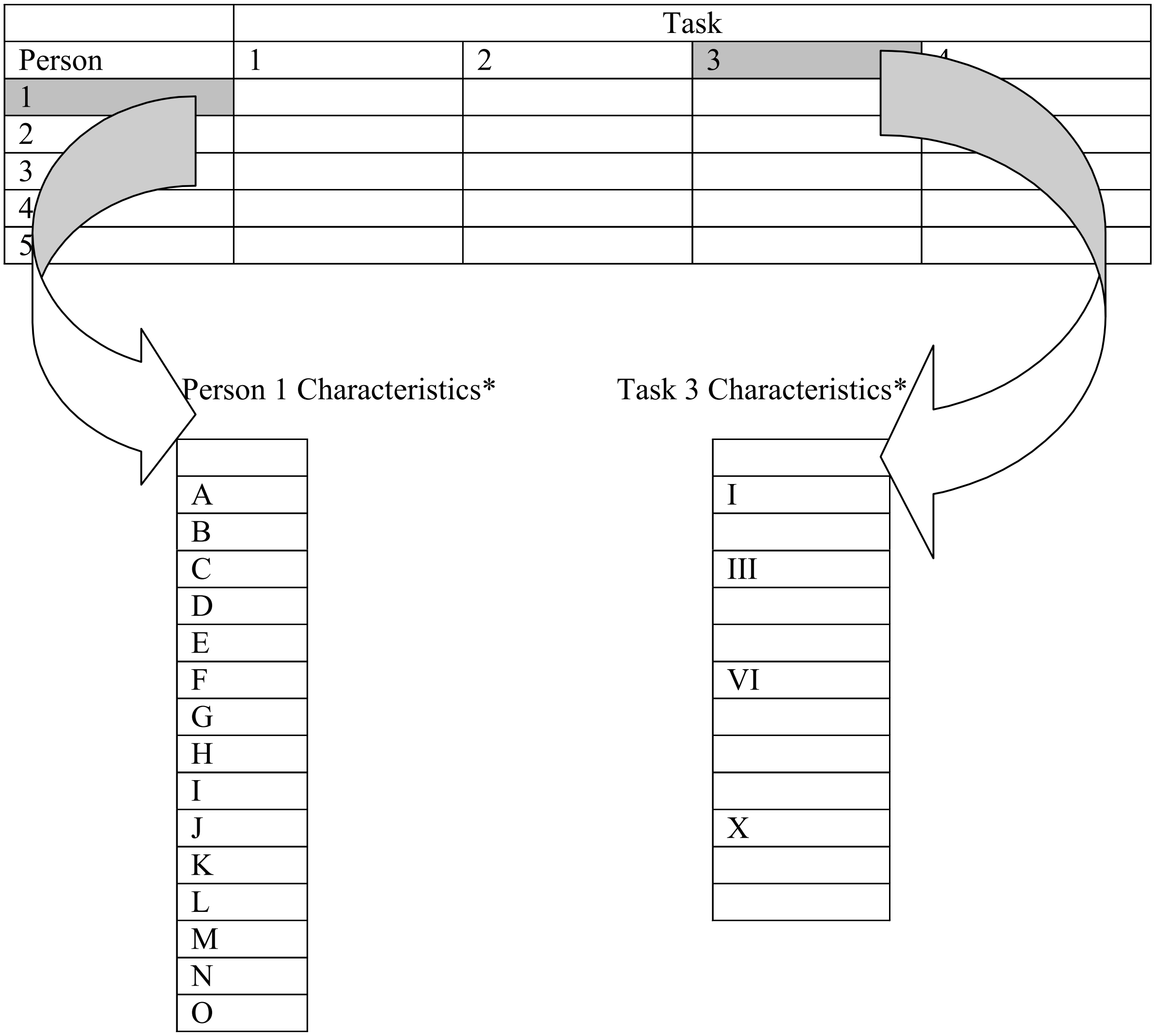
Conceptual Approach

Our area of study is the encounter of individual test takers with test tasks, investigating under what conditions target skills are properly conveyed in this interaction and, in particular, when communication about targeted information becomes systematically contaminated, misunderstood, or distorted. We define this contamination or distortion as error which occurs in a regular and predictable way when individuals with specific characteristics interact with specific task factors, and error which influences task performance but is not part of what one intends to measure. For some students this influence results in a significant source of interference in accurately measuring target knowledge and skills and in interpreting scores. For the moment we are calling the interference an issue of person/task *access*. Our focus within the encounter revolves around how the same assessment tasks differ from individual to individual with respect to interactive processes of access a) to the task meaning, b) to the individual's problem solving mechanisms (eg. identifying problem solving strategies, selecting and assembling the proper strategies, and implementing the strategies to come to a solution), and c) to retrieving, organizing, and demonstrating the task solution.

We begin this inquiry by outlining in general the person/task encounter components. Individual test takers bring a collection of characteristics that vary over individuals (*individual difference variables*), such as target construct knowledge, problem solving strategies, perception and attending skills, schooling experiences and various kinds of non-target knowledge and capabilities. Each assessment item has a set of construct relevant and irrelevant characteristics that also vary over tasks (*task difference variables*). These include characteristics such as a description of requested target knowledge and skills, procedural requirements, contextual variables, a way of communicating what is required or how results should be conveyed, and contextual or associated tools to use in problem solving. For any person characteristic there is a subset of item characteristics that are relevant, either primarily or secondarily. Likewise, for any item characteristic there is a subset of person characteristics that are primarily or secondarily relevant and that interrelate with the item characteristic to impact performance on that particular item at some level. Finally, the encounter itself involves processes. Cognitive psychologists have outlined problem solving processes that occur when students attend to task requirements (e.g. see Pelligrino, Chudowski, & Glaser, 2001). While individual characteristics vary over persons, and item characteristics vary over tasks, the encounter processes, including both targeted processes and access processes, vary over both person and task simultaneously.

Step 1

Person by Task Interaction



	Task			
Person	1	2	3	4
1				
2				
3				
4				
5				

Person 1 Characteristics*

A
B
C
D
E
F
G
H
I
J
K
L
M
N
O

Task 3 Characteristics*

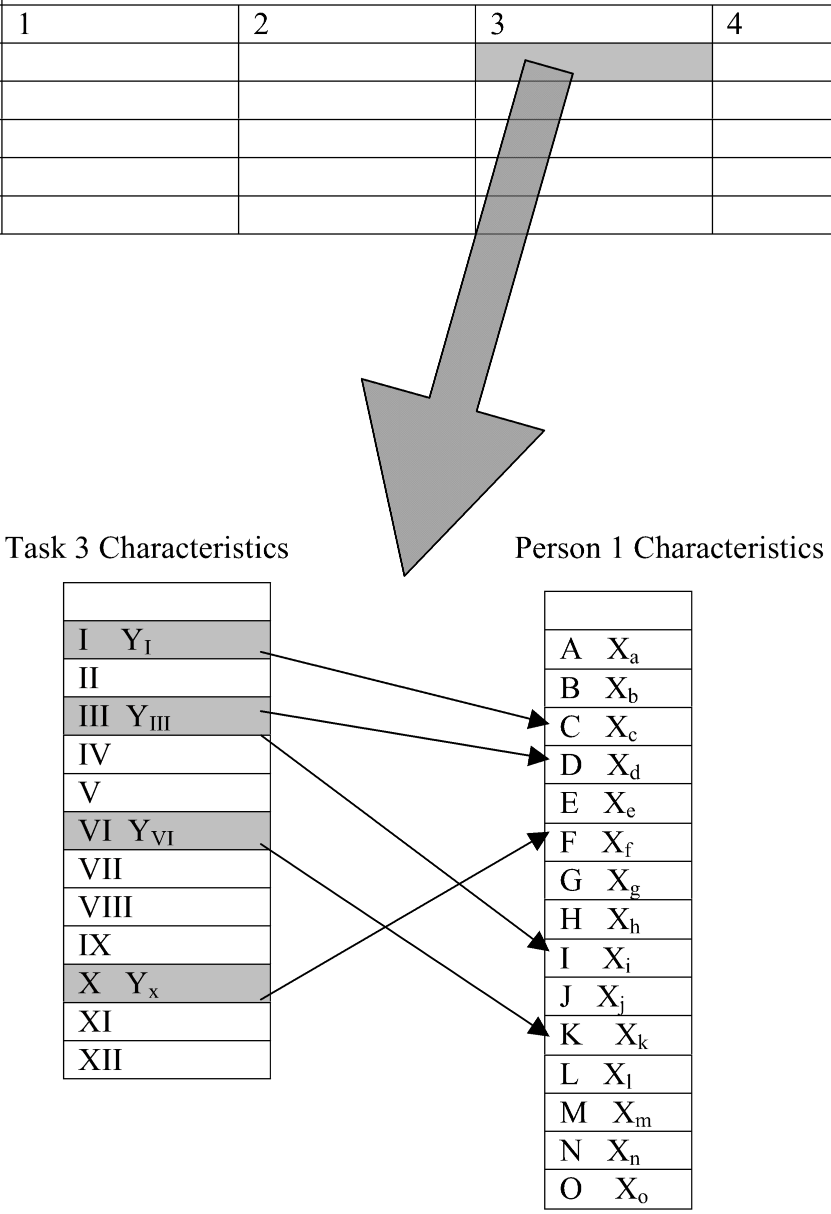
I
III
VI
X

*Person and task characteristics may be continuous, discrete, or dichotomous.

Step 2

Interaction Space between Person 1 and Task 3

	Task			
Person	1	2	3	4
1			▨	
2				
3				
4				
5				



Step 3

Match between Task 3 Characteristics and Person 1 Characteristics

Task 3				
Person 1	I	III	VI	X
A				
B				
C	-			
D		+		
E				
F				-
G				
H				
I		+		
J				
K			+	
L				
M				
N				
O				

+ indicates a sufficient match between task characteristic and person characteristic

- indicates an insufficient match between task characteristic and person characteristic

Example of Compensatory Rule:

If I/C interaction = - and III/D interaction = +, then I/C interaction = +.

Example of Conjunctive Rule:

The VI/K interaction must be positive (necessary although not sufficient) for the targeted interaction to be unaffected.

Our approach is to identify person/task encounter processes and explain and model them in terms of individual difference and tasks difference variables. To do this we observe and investigate why students respond to tasks as they do. Concurrently we are iteratively formulating mathematical/statistical functions to model these phenomena by treating person-task process characteristics as outcome variables and individual and task differences as explanatory variables. Because person and item characteristics are classified as target or irrelevant, the impact of the specific irrelevant person/task access interactions can be viewed.

We observe that the person/task access interactions occur primarily at three points:

1) For the interaction between the student's targeted knowledge and skills and the item's request for targeted information to successfully commence, the student must have sufficient access to how the meaning and the requirements are conveyed in the task.
2) For the student to initiate and sustain problem solving activities relative to the task requirements, students must be able to access their procedural skills and other ancillary content, and have the tools necessary to implement the activities. (The ancillary content and skills would be considered construct irrelevant and within the domain of access to the extent that they are not part of the target requirements.)
3) For the student to represent their solution to the task requirements, students must be able to access their representation skills commensurate with the representation constraints in the particular task.

As we view and model the impact of specific person/task interactions and their effects on performance, we also attempt to identify, build and investigate demonstration test development procedures and products that might be useful in addressing access by ameliorating the overall interference of the irrelevant interactions without adding significant additional sources of error to student scores. The argument is that these procedures and products could be designed to produce common inferences, and, where possible, comparable scores over testing components which vary in appropriate ways over test takers. The grounds for common test inferences are traditionally found in a *procedural argument*: common content in items and a common approach for synthesizing and summarizing items and response data over items. The latter part of this procedural argument required standardized conditions of observation as a key aspect of synthesizing item data. However, based on developments in fields such as those identified 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, the properties of tasks or items designed to elicit the observations, student characteristics that impact testing and items, necessary observations, and the assessment situations where students interact with assessment requests. This approach suggests that data may be collected under alternate conditions (Mislevy, 1995). By minimizing the influence of irrelevant input on student performance without adding significant additional sources of error we increase the validity of the task inference without sacrificing reliability. At its crux, Kopriva (1999; 2001) suggests that, when common inferences from a robust conceptual argument are applied to assessment results for which there is sufficient evidence of minimizing systematic error across testing procedures and products based on the arguments explained in this paper, this should provide the grounds and be the basis for determining the validity and comparability of scores.

Underpinnings to Models

In the field of assessment, models are generally fixed and inferences are formed that are conditional upon those models. Here, we seek to generate evidence about processes and in doing so create evidentiary chains that lead us to new models.

We begin with established models in order to explain some of the components of the new chains, variables, and models.

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

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

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

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

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

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

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

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

$$P(X_{is} = 1 \mid \underline{\theta}_s, \beta) = \frac{\exp(\sum_m \theta_{sm} - \beta_i)}{1 + \exp(\sum_m \theta_{sm} - \beta_i)} \qquad (2)$$

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

$\beta_i =$ the difficulty of item i

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

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

For this area of study, we propose using the compensatory approach to model access, while also attempting to capture the conjunctive relationship between ancillary and targeted knowledge for both student and item factors and perhaps within item and student factors. We expect that these relationships can be explained and be modeled. For instance, the multidimensional access process would capture the compensatory and non-compensatory factors jointly as a more general interactive function,

$$P(X=1|\Theta, A) = F(A, \Theta) \qquad (3),$$

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

In one example of where we began our thinking, we discussed building the compensatory element into the conjunctive model given in equation (1) to yield the following:

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

However, we are concerned that the relationship isn't consistently as straightforward as equation 4 would suggest. Rather, we believe that under specific cases or types of item-student and/or access-construct interactions, the relationships will reflect a weighted and more nuanced interdependent approach to student performance. For example, items might have the same vectors of item factor levels, yet the required student factor levels for successful performance may differ, depending on the joint status of the factors and other characteristics of the items.

Modeling the item to student access matches

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

> 1) the identification of the target and ancillary factors evoked by particular items across students to form a target and access profile vector for each item
> 2) the identification of interactive rules within the item profile vector for specified factors across students with different student profiles
> 3) the identification of factors in student profiles across items to form a vector profile by student

4) the weighting and interaction of target and ancillary factors within student vectors to provide prioritized matching expectations that differ by student.

Identification of target and non-target factors and interactive rules for persons and items includes determining the presence or absence of specific components in items or in students, and some indication that quantifies the threshold amounts of each component. Interactive item rules explain when the item requirements or problem solving mechanisms can be accessed by more than one factor or when the presence of one factor could lower the threshold for another factor. For students, some factors may be necessary (although not sufficient), or certain alternative student profiles may apply to specific item configurations

Preliminary work in modeling these matching procedures and rules has suggested that they can be modeled and programmed--for students with different profile vectors and targeted knowledge levels, and for items and item types that measure various types of targeted information and evoke different access needs. In general the approach being taken is to create a target/access profile vector for each student that will be used across items, and an item target/access vector for each item that explains what factors are evoked in that item. In each case, as noted above, the student and item vectors are "customized" by additional rules of compensatory and conjunctive interaction and weighting that differ over students and items, respectively. Once these profiles are identified and the rules programmed by student or item, it appears that the matching can be completed electronically. A best match among like items can be found for each student for each identified targeted construct element that a test is intending to measure.

Ultimately, we expect to include factors for each point of the problem solving process into the item vectors and possibly the student profiles. Because only a restricted set of factors will be evoked at any one point in the process, it is expected that this refinement will add an additional set of rules to the established vectors, but not be prohibitive.

Implications for test development

As some variations are made to particular items or the methods in which they are administered, the approach set forth here would need to address how this can be done in a technically defensible manner, so that the integrity of the target is maintained, and adequate reliability and comparability of scores are retained over variations. We believe that the *essential* focus in permitting and constraining what variations would be permitted is that target-invariance over item variations must be preserved. We believe that this criterion becomes the benchmark for improving overall validity in testing. It also becomes the criterion for ensuring and maintaining reasonable validity across students and across item variations within a testing system. Further, we believe that when there is appropriate evidence about the target-invariance of items, scores from the variations should be seen as comparable.

While this is a different way of thinking about validity and comparability, it is not without precedent. One example is the use of scoring rubrics in large scale testing which, with appropriate technical and oversight supports, standardize the constructs and evaluation processes while allowing specific responses to vary from student to student. Another example is computer adaptive testing, which provides different combinations of items to each student, yet yield measures of ability that are comparable. Certainly, such a variation-inclusive system would need to address how and when additional sources of error due to the variations would offset errors in traditional tests that occur from decreased validity of score inferences for specific students.

Therefore, one of the primary implications of this approach for test development becomes developing item templates, procedures and criteria that defensibly address how to identify and precisely define what is the targeted knowledge in items, and identify and define what are the construct-irrelevant aspects in particular items (in other words, identify what is being measured by specific items that is NOT part of what the target). Then, any item variations that are produced should differ *only* in the type and degree of ancillary construct irrelevant factors they evoke. This is particularly of interest for large scale testing purposes, where the technical rigor is the most salient. This approach suggests addressing:

- the identification of an item development framework that specifies how to design target-invariant items in a large-scale setting
- the identification of the item and test development process and set of procedures to produce items and tests within this framework
- the development of the first set of model items and testlets that fit these specifications
- an initial empirical investigation into the validity of this approach.

We believe that this approach would produce templates that are more constrained than many envision when they discuss the use of templates or item shells in testing (e.g. Haertel and Mislevy, 2001; Solano-Flores, 2002). For instance, in constraining the template to keep salient contextual data as well as target specifications constant, the concept minimizes error due to many of the additional sources of variation in items. Such sources of variation would be those that occur across items that are derived from templates whose criteria include only the content objectives of items. We envision that such an approach would allow for the construction of a set of items that are virtually interchangeable, while also meeting the needs of diverse students in a more appropriate way. It marks an intermediate step between one item and a great many possible items that share specified general template criteria.

Investigations underway to date

In beginning to investigate the concepts set for in this overview, we have three studies underway. In all three, the target groups are English language learners and poor readers, who we know traditionally are not well served by many of our tests. Work with these groups and controls should provide a base for beginning to identify salient item and student characteristics, build models, study the impact of construct irrelevant interference, and develop procedures and products to ameliorate significant problems in the person/task interaction space.