[to appear in *Language Assessment*]

Design and Analysis in Task-Based Language Assessment

Robert J. Mislevy[1], Linda S. Steinberg[2], and Russell G. Almond[2]

September, 2001

[1] University of Maryland

[2] Educational Testing Service

Acknowledgements

We are grateful to John Norris and Lyle Bachman for comments on earlier versions of this paper, and to the staff and consultants of TOEFL for many stimulating and enlightening conversations over the years. The first author received support under the Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U. S. Department of Education. The findings and opinions expressed in this report do not reflect the positions or policies of the National Academy of Sciences, the National Research Council, the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U. S. Department of Education.

Design and Analysis in Task-Based Language Assessment

Abstract

Task-based language assessment (TBLA) grows from the observation that mastering the grammar and lexicon of a language is not sufficient for using a language to achieve ends in social situations. Language use is observed in settings that are more realistic and complex than in discrete skills assessments, and typically require the integration of topical, social, and/or pragmatic knowledge along with knowledge of the formal elements of language. But designing an assessment is not accomplished simply by determining the settings in which performance will be observed. TBLA raises questions of just how to design complex tasks, evaluate students' performances, and draw valid conclusions therefrom. This paper examines these challenges from the perspective of "evidence-centered assessment design." The main building blocks of are student-, evidence, and task models, with tasks to be administered in accordance with an assembly model. We describe these models, show how they are linked and assembled to frame an assessment argument, and illustrate points with examples from task-based language assessment.

Design and Analysis in Task-Based Language Assessment

Introduction

There has been increasing interest in the language testing community in Task-Based Language Assessment[1] (TBLA), or "the process of evaluating, in relation to a set of explicitly stated criteria, the quality of the communicative performances elicited from learners as part of goal-directed, meaning-focused language use requiring the integration of skills and knowledge" (Brindley, 1994, p. 74). Interest in TBLA can be attributed such factors as the alignment of task-based assessment with task-based instruction, positive "washback" effects of assessment practices on instruction, and the limitations of discrete-skills assessments, or DSAs (Long & Norris, 2000). DSAs focus on the knowledge of language per se, exercising points of lexicon, syntax, and comprehension with discrete and largely decontextualized test items. Recognizing the fact that knowledge of vocabulary and grammar (linguistic competence) is not sufficient to use a language to achieve ends in social situations, TBLA embraces a broader conception of communicative competence. In addition to linguistic competence, consideration broadens to the social context of language use (sociolinguistic competence), pragmatic considerations in using language to achieve goals (strategic competence), and familiarity with forms, customs, and standards of communication above the level of sentences (discourse competence).

While we have models of the range of characteristics that make up student L2 competence (e.g., Bachman & Palmer, 1996, Chap. 4), what has been lacking is a systematic means for designing performance assessments that will directly and adequately inform the particular kinds and qualities of inferences that need to be made for various assessment purposes, such as program accountability and evaluation, summary evaluation of students' proficiencies, evaluation of students' progress on what they have been working on, predictions of success in particular language use settings, and needs assessments for guiding instruction. In particular, a systematic approach to assessment

[1] Also known as Task-Centered Assessment (TCA) and Task-Based Language Testing (TBLT). We use the terms as synonyms.

design would address the following questions that confront current notions of task-based language performance assessment:

- What does a given performance task measure?
- How do you score tasks?
- What does a collection of tasks measure?
- What features of tasks determine their difficulty?
- What features of TBLAs determine their reliability?
- What factors affect the validity of TBLA?
- What factors affect its generalizability?

The challenges of complex performance-based assessments are not unique to language testing, of course. Similar motivations, demands for authenticity, and similar difficulties with design and application are the topic of much discussion in educational measurement more generally (e.g., Wiggins, 1993; Wolf et al., 1991). This paper examines the design of complex assessments from the perspective of "evidence-centered design" (ECD; Almond et al, 2001; Frase et al., in press; Mislevy, Steinberg & Almond, in press; Mislevy, Steinberg, Breyer, Almond, & Johnson, 1999, in press; Mislevy, Steinberg, Almond, Haertel, & Penuel, in press). In this paper we present the rationale for ECD, review the high-level ECD models, then discuss issues design and analysis of TBLA from its perspective.

The Rationale for a Design Framework

Problems of design and analysis are straightforward in discrete skills assessments, because it is possible to focus an item's demands on particular points of lexicon, syntax, morphology and so on, typically delivered in a receptive mode. DSA "works" in an important sense: Now-familiar procedures have evolved for designing items, evaluating responses, and accumulating the information over items. There is a methodology in place for building a coherent evidentiary argument from what a student says or do in the assessment to what the student knows or can do—but with direct evidence about only these very focused and limited uses of language. The problem is, of course, that the

inference of interest may call for a broader range of knowledge, and the ability to put it to use, than these kinds of items can reveal.

The concern in TBLA extends beyond knowledge of language per se, to the ability to deploy language knowledge appropriately and effectively in educationally or professionally important language-use settings. Whatever topical, social, and pragmatic knowledge these situations demand, of native as well as of second-language speakers, must be considered. It is no longer so clear how to construct tasks that elicit desired L2 performances, which aspects of performance to evaluate, how to integrate information from multiple tasks, or what kinds of inferences to draw about students. For most inferential purposes, it isn't enough to simply create task situations that seem important in and of themselves or demand competences of interest (Messick, 1994). Understanding what features of tasks influence their difficulty is a good next step, but it isn't enough either. Developing psychometric models that deal with more complex data may be necessary too, but it still isn't sufficient. None of these lines of work, by themselves, will produce a coherent evidentiary argument. We desire a framework that integrates all of these elements from the very beginning, from an assessment's purpose to inferences about students.

There are practical benefits to explicating an assessment argument, to be sure. Having laid out at a higher level of generality the aspects of competence or capabilities we are interested in, schemas for eliciting them, and rubrics for characterizing the relevant features of performances, we can create a continuing series of tasks and know 'how to score' each one. But more important is the foundation for validity and generalizability. Generalization depends on systematic thinking about features of tasks and their connections with students' competences and capabilities, connections observed in target language use situations and continually sharpened and delineated by validity research. Messick (1994) argues that this is the only way we can determine whether a set of tasks fails to provide evidence about what we care about ("construct underrepresentation") and the degree to which tasks demand knowledge and skill that aren't central to our purposes ("construct irrelevant variance").

This reasoning holds, we believe, for more complex assessments in any substantive domain. Generally applicable principles of evidentiary reasoning help us structure coherent assessment arguments, while domain-relevant knowledge provides for the content of the argument; that is, what we want to say about students, and what kind of evidence we need to see. In language assessment, the principles of assessment design must be applied in concert with principles of communicative competence. We need to build an assessment argument around what we are learning about the ways that language knowledge interacts with other knowledge to constitute ability for use in target language use settings (McNamara, 1996, p. 48). Evidence-centered design offers a framework for first working through the structure of the argument, then designing elements of that can be assembled to transform that argument into an operational assessment.

The ECD Framework

This section briefly introduces the basic high-level models of the evidence-centered design framework. Then we'll look more closely at each. Along the way we'll discuss issues that concern relationships among the models.

Figure 1 is a high-level schematic of four basic models in what we call a "conceptual assessment framework," structures we suggest must be present (at least implicitly), and must be coordinated (at least functionally), to achieve a coherent assessment. They are *student-*[2], *evidence-*, and *task-models*, and an *assembly model* that governs the way that individual tasks are brought together to form assessments.

[[Figure 1 – ECD models, including assembly model]]

In brief, the *student model* specifies the variables in terms of which we wish to characterize students. *Task models* are schemas for ways to get data that provide evidence about students. *Evidence models* consist of two components, which are links in the chain of reasoning from students' performances to their knowledge and skill: The *scoring component* contains procedures for extracting the salient features of student's performances in individual task situations—i.e., ascertaining the values of observable

variables—and the *measurement component* contains machinery for updating beliefs about student-model variables in light of this information. An operational assessment will generally have one student model, but may use multiple task and evidence models to provide data that provide evidence about different aspects of skill and knowledge. An *assembly model* specifies how individual tasks are combined to produce an assessment. Together, they make explicit the evidentiary argument that underlies an assessment.

A quote from Messick (1994) matches up fairly well with these models. Messick's comments are phrased in terms of "construct-centered" assessment design, which some writers such as Frederiksen and Collins (1989), see as antithetical to "task-based" assessment. It will become apparent why we don't think so.

> A construct-centered approach would begin by asking what complex of knowledge, skills, or other attribute should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are otherwise valued by society. Next, what behaviors or performances should reveal those constructs, and what tasks or situations should elicit those behaviors? Thus, the nature of the construct guides the selection or construction of relevant tasks as well as the rational development of construct-based scoring criteria and rubrics. (p. 17)

The Student Model

"What complex of knowledge, skills, or other attributes should be assessed?" The student model in Figure 1 depicts student model variables as circles. Configurations of their values are meant to approximate selected aspects of the infinite configurations of skill and knowledge real students have, as seen from some perspective about skill and knowledge in the domain. These are the elements we use to accumulate evidence about students' competence, as we seek to fulfill the purpose of the assessment. We don't see the values of SM variables directly. We see what the students say or do, and construe that as evidence about SM variables.

---

[2] We use the term 'student model' to be consistent with our other publications, in which most of the applications are educational. 'Examinee model' might better communicate the general nature of this model.

We use probability-based reasoning to manage knowledge about a given student's unobservable values for SM variables at any point in time, expressing it as a probability distribution that can be updated in light of new evidence (Mislevy & Gitomer, 1996). Psychometric models such as classical test theory (CTT), item response theory (IRT), factor analysis, and latent class models are special cases of this approach.

Clarifying the Meanings of "Student Model"

We need to distinguish two ways the phrase "student model" is used in the literature. The student model we are depicting in Figure 1, and the way we will be using the term, refers to a piece of machinery—a set of variables in a probability model, for accumulating evidence, for backing claims. We may have an intended interpretation for the variables, but its meaning is acquired through its correspondence to what we can observe. Operationally, SM variables are summaries of patterns of values of observable variables, along lines we build into the evidence-model structures (discussed below), as evoked in part by features we build into tasks (also discussed below).

By way of contrast, Bachman's model is a student model in a different sense, a substantive sense, in this case addressing components of competence. Other substantive models address how performance is produced. Substantive student models inform our thinking about the claims we might want to make about students, but they cannot determine the statistical student model that is appropriate for a given assessment. The purpose of the assessment further determines the focus and the grainsize of the variables in the statistical student model.

The point here is the inescapable needs to spell out the purpose of an assessment, and to then determine the aspects of language use proficiency that serve that purpose, be they phrased in terms of competences or in terms of behavioral tendencies in situations with given features. Target language use (TLU), to borrow a phrase from Bachman and Palmer (1996), may be the starting point, but they are not the ending point. They can help us determine the kinds of claims we want to make about students, and from their guide our construction of the pieces of machinery we need to build an argument to back

those claims: statistical student models, measurement models, evaluation rules, and generative task models.

The Relationship between Claims and Student-Model Variables

Claims in assessment are statements we'd like to make about what students know, can do, or have accomplished, explicitly or implicitly in relation to contexts. They may be more or less detailed, and they can range from everyday language to jargon, but in any case they are substantively meaningful statements about students and subject matter.

McNamara (1996) discusses a language assessment for medical students for whom English is a second language. Treating and counseling English-speaking patients is the TLU, and having students work through consultation scenarios with English speaking simulated patients provides direct evidence about their capabilities to do this, in way that confounds language use with medical knowledge and interpersonal skills. A hiring officer would address a global claim about how successfully a student can carry out interactions in English with the kinds of patients and problems that are routinely encountered her clinic. A school advisor would need finer-grained claims, because simply learning that a student could not handle consultation in English satisfactorily would not indicate whether the student's difficulties lie with language use per se, or with medical knowledge or interpersonal skills. Her claims would be at a level that would guide placement into practice sessions.

In contrast to claims, SM variables are pieces of machinery. A student model is a mathematical structure containing variables and a joint probability distribution for them. Within this structure, the calculus of probability specifies how new information about any set of variables changes the probability distribution for the rest. This machinery does not know or care about how we interpret the variables, as characteristics of students or summaries of performances. But it is the machinery we use to characterize the evidence we have to back substantive claims. Let us consider four approaches an assessment designer can relate claims and SM variables.

One possibility is one-to-one relationship between a claim and a continuous SM variable. The claim is whether a student 'has mastered' some skill or knowledge, defined

as a sufficiently high probability of success on tasks in a specified domain of tasks. The tasks may be simple or complex, but each yields a single dichotomous observable variable.[3] Here the SM variable is interpreted as a student's propensity to make performances at some level, perhaps as rated at several levels of quality. Observing performance on each task adds another nugget of evidence. In this approach the statistical model can be used to back statements such as "The student's probability of a correct response is at least 80%", or "The student's probability is above 50%, so he doesn't need to go back to lesson 1, but it's below 75%, so he ought to work through the practice drill." The motivating claim is a statement that anchors the upper end of a continuous proficiency variable, and lower values of the variable correspond to lesser degrees of proficiency.

A second approach encompasses multiple claims and a single SM variable with a finite number of levels. Each value of the SM variable matches up one-to-one with a particular claim or set of claims, as in the American Council on the Teaching of Foreign Languages guidelines for language proficiency (ACTFL, 1989, 1999). There are ordered levels in separate scales for Reading, Writing, and Listening, which each have nine levels, and Speaking, which has ten levels in the 1999 revision. The ACTFL guidelines move from Low Novice, in which the student can typically only use language in a given modality in rudimentary ways, up through Superior. Each level is described in terms of several kinds of things a typical student at that level can do, in situations with certain key features (see Table 1), each of which is a claim in its own right. The intent is that statements within a level go together well enough to characterize a student in terms of a single level, although there will be some performances above or below that level—a cost of supporting many distinct claims with a single SM variable. While the degree to which this intent is realized can be debated (e.g., Bachman & Savignon, 1986), partitioning language use into four separate modalities precludes SM variables based on the ACTFL scales from directly addressing claims that would address the joint use of two or more modalities.

[3] The more complex a collection of tasks is and the more they differ from one another as to the mix of competences they demand, the lower the reliability will be of the score; i.e., the less information about the

[[Table 1: ACTFL scale excerpts]]

A third alternative for addressing multiple claims with a single SM variable is to model probabilities of given responses in settings with particular key features. In contrast to the approach described above, a claim is not associated with a single value of an SM variable. An example is the Document Literacy scale from the Young Adult Literacy Survey (YALS; Kirsch & Jungeblut, 1986), in which (a) tasks requiring very simple responses are generated in accordance with numbers or degrees of features that are salient under their cognitive model, (b) a single IRT variable θ accounts well for performance across all the tasks, and (c) a regression model using salient features as predictors accounts well for task difficulties. Salient features include the kind, structure, and complexity of the documents at issue, and the complexity of what the student is asked to do in terms of factors such as number of steps and congruence with the document. One can generate a whole family of claims using these features; for example: "Eighty-percent of the time, the student can do **feature matching** in documents that are **structured according those features**, when **three** features need to be matched and there are **no close distractors** in the text." The highlighted phrases are specific values for slots that can be filled in with values of TM variables to produce a large number of claims. That is, substantive meaning is imparted to the ITR-based SM variable θ by describing the likelihood of success of students at any given level in situations described by specified levels of TM variables. Because of properties (b) and (c), knowledge about a student's θ in terms of a current distribution can be translated to support for any claim that can be generated in this way.

A fourth approach directly tackles the interactions among competences and contexts: multiple SM variables are called upon to express evidence for a claim. This is appropriate when multiple aspects of knowledge or skill are required in combination to support a claim, and students exhibit different patterns of proficiency. Distinct SM variables are used to maintain belief about distinct knowledge/skills, and a claim is associated with particular patterns across them as they are called upon in settings that

---

overall proficiency SM variable will be conveyed by each task's observable variable, regardless of how rich and fertile the observation setting may be.

stress or combine competences in different combinations. Thus, students' proficiency in a domain can be described in terms of which skills they possess (via SM variables), tasks can be described in terms of which skills they require (via TM variables), and the outcomes expected from any particular matchup can be described in terms of values of observable variables. This is a multivariate generalization of Approach #3 above. A claim can be stated about a student's proficiency with respect to a class of tasks with the same salient features. The evidence for such a claim is contained in the SM as the joint distribution for the particular skills, in the particular combinations, that are called for by tasks with these features.

For an example of this multivariate approach, we will use the Document Literacy scale defined above and introduce another scale concerning Written Explanation. We saw above how the a student's reasoning and receptive capabilities with documents can be linked with the features of documents. We can now generate a family of tasks that (a) provide documents of varying complexities and (b) require written responses of varying complexities. For example, we can have simple documents and ask for simple phrase responses. We can provide simple documents but require elaborated written responses. We can provide complex documents and directives yet ask for simple responses, or provide complex documents and demand elaborated responses. In the section on measurement models we will discuss how to sort out evidence about reading and writing competencies from multi-skill tasks such as these.

Complex performance situations can be difficult for different reasons, relating to different aspects of competence. If the claims we want to make require sorting out the reasons that different people fare well or poorly in settings with different features, we need a student model with variables that can make the necessary distinctions (specifically, multivariate models, as we will discuss below). We also need to identify features of settings that stress different aspects of competence, and know how to sort out the evidence about the different aspects of competence in these complex situations. We say a bit more about these issues in the following sections on evidence and task models, and return to them again in terms of the interplay among student, evidence, task, and assembly models.

Evidence Models

"What behaviors or performances should reveal those constructs," and what is the connection? Evidence models lay out the argument about why and how observations in a given task situation constitute evidence about student model variables. A given assessment could use several evidence models, if it uses a number of qualitatively different schemas for capturing and evaluating evidence. Figure 1 shows there are two parts to an evidence model, the evaluation component and the measurement component.

The Evaluation Component (Task-Level Scoring)

The *evaluation component* concerns extracting the salient features of whatever the student says, does, or creates in the task situation, or the *work product*. A work product, represented in Figure 1 by the rectangle containing a jumble of shapes at the right, is a unique human production—a mark on an answer sheet, written directions from the hotel to the bank, a sequence of utterances in a conversation with an interviewer about the weather. One task can yield multiple work products.

The squares coming out of the work product represent *observable variables*, or summaries of what the assessment designer has determined to be the key aspects of the performance. The evaluation rules indicate how to carry out these mappings from unique human actions into a common interpretative framework. Evaluation rules for complex performances, rubrics for rating scales in particular, represent choices about what is valued and how it is to be evaluated. Aspects of both the product of a performance and of the performance itself are possible to evaluate, and multiple evaluations of either or both can be made, especially if they bear on different aspects of competence.

As an example, consider a learner's oral directions to the library. We might rate its overall effectiveness, but we could further distinguish the aptness of its content, its complexity, the accuracy of its grammar and vocabulary, and its sociolinguistic appropriateness. An important insight from psycholinguistics is that these latter qualities tend to trade off against one another in use, as people decide how to expend their limited cognitive resources to get the most benefit from the competences they possess (Skehan, 1998, pp. 167ff). What we should capture with evaluation rules, therefore, depends on

which aspects of competence, usage, and effectiveness we need to serve the assessment's purpose.[4] The overall evaluation could suffice for predicting effectiveness in a job setting, the key features of which appear in the task. Feedback on particular competences that are required in the response would require ratings that captured evidence about aspects of the performance that reveal them (perhaps confounded with other competences). Knowing about tradeoffs argues for modeling these multiple ratings as conditionally dependent (Bradlow, Wainer, & Wang, 1999).

Determining evaluation rules is thus a judgment that we can not expect experts to agree on, if they are provided tasks and sample performances but not assessment purposes. In particular, we cannot design task situations and capture work products, then hope some else can tell us how to evaluate the performances (Brindley, 1994, p.79).

A First Aspect of Reliability

Two recurring issues in TBLA are cost and reliability. Reliability is central to our argumentation theme. Cost is not, but it deserves mention because it regularly trades off against reliability and validity. Having thought through at a higher level of generality what kinds of behaviors we need to see, in situations with what kinds of features, we can then explore a range of possible ways to make and evaluate observations. Each way of getting evidence will have its own configuration of required skills; some are relevant and others are not, although we can't know which is which until we specify the intent of measurement. The quality of the mapping from the performance to the observable variables is a first aspect of reliability.

Consider, for example, the issue of co-construction of meaning in conversations. To get direct evidence about a student's capability to do this in given situations, we need to observe her reacting to a conversational partner, and adding to the interaction in a way that would add to their joint understanding. Interviews are the usual way this evidence is gathered, but interviews are costly. Simpler interactions can be provoked with computer programs; this is done routinely in language instruction programs such as *Herr Commissar* (DeSmedt, 1995). The costs are lower, but so is fidelity. The computer

[4] It is an important consideration of fairness, not to mention validity, that students know what this is, for it

presentation both misses some aspects of conversation with a real person (e.g., a more constrained range of understanding) and introduces some irrelevant skills (e.g., interacting with whatever interface is required). Recent work on the Computerized Oral Proficiency Interview (Kenyon & Malabonga., 2001) explores a configuration that lies between these extremes. Whether the resulting tradeoffs among evidentiary value, cost, and consequences is favorable must be examined case by case. We can expect that accumulating experience will provide us with forms and configurations that work well in given settings for given purposes.

Regarding the reliability of rating schemes and rating procedures, we will just note two fronts on which considerable improvement has been made. The first concerns technology for recording and transmitting performances (Sheingold & Frederiksen, 1994). Formerly ephemeral performances can now be replayed at will, so the meaning of evaluation rules can be disseminated more widely and consistently to raters, instructors, and students. The second concerns psychometric models that incorporate rater effects. Although rating is formally part of the evaluation component, we will say more about it in the following section on measurement since these analyses typically address measurement and scoring jointly.

The Measurement Component (Test-Level Scoring)

The *measurement component* of an evidence model expresses how the observable variables depend, in probability, on SM variables. This is the embodiment in machinery of another part of the evidentiary argument: how to synthesize evidence across multiple tasks and different performances. We see in Figure 1 that the observable variables are posited to depend on some designated subset of student model variables. Two familiar examples are IRT and factor analysis. In IRT, there is only one SM variable, θ, and the measurement component concerns the tendency to make a higher quality response rather than a lower quality one given θ. In factor analysis, the a multivariate student model is to be discovered from patterns in the data, and the measurement model for an item is specified by its factor loadings.

---

will surely influence their performances.

An approach of particular relevance to TBLA uses a multivariate student model and tasks designed to elicit evidence about particular SM variables through particular choices of task features. The structure of the measurement component is predetermined by the structures built into the tasks. Conformable measurement models can include item-level confirmatory factor analysis (e.g., Muthen, 1988), structured IRT models (e.g., Adams et al., 1997; Embretson, 1998), and modular assemblies of Bayes net fragments (e.g., Almond & Mislevy, 1999; Mislevy, Steinberg, Breyer, Almond, & Johnson, in press).

Embretson's (1985) edited volume was a watershed event in the emerging confluence of psychometric modeling, substantive theory, and task design. Here are the key ideas: The features of a task can be used to direct the evidentiary focus of tasks on aspects of competence or proficiency, and mediate the stress put on those aspects. These relationships can be built into statistical models such as those cited above to (a) make explicit the expected relationships, (b) guide task construction, and (c) exploit the structure in inferences about SM variables. Applied to TBLA, we want to integrate into the measurement model what research reveals about the relationships among aspects of the ability to use language and features of performance situations (e.g., Brindley, 2000; Brown, Hudson, Norris, & Bonk, in press; Norris, Brown, Hudson, & Yoshioka, 1998; Skehan, 1998). When tasks are designed around schemas for which conformable measurement model structures have been provided, we know ahead of time how to sort out evidence about complex student models from complex performances.

For an illustration, recall the reading/writing tasks we introduced to discuss the relationship between claims and SM variables. A student must read a passage then produce a written response in accordance with a directive. A student's response to Task j is rated with respect to three qualities: X_{j1} is language usage without respect to its substance, X_{j2} is the appropriateness of the substance, and X_{j3} is the overall effectiveness of the response. We have two SM variables, *Understanding Documents* and *Writing*; call them θ_U and θ_W. Features of the document, the directive, and their interrelationships—collectively denoted Y_j for Task j—influence how much demand is placed on θ_U and θ_W with regard to each observable variable. Features of documents and directives, it will be recalled, concern type and complexity of the document and task complexity; we'll denote

the subset of features that pertain specifically to reasoning with the document as Y_{jU}. Features of the task that pertain to the required response indicate whether the student must produce a simple phrase in response, a well-formed sentence, a paragraph, or a structured multi-paragraph explanation; we'll denote these as Y_{jW}. Other features of the task affect performance globally, such as time limit; we'll denote these as Y_{j+}. We could generate any number of specific tasks from this same task model; each would have its own documents and directives and therefore its own Y_js that determine the load on *Understanding Documents* and *Writing*. As in the models of Fischer (1973), Embretson (1985), and Adams et al. (1997), the probabilities of response outcomes will be modeled as probabilistic functions of the relevant features of the tasks.

Figure 2 depicts the measurement model for a single task of this type. The circles are the SM variables, the squares are the observables, and the matrices represent conditional probability distributions. Appropriate subsets of the Y_js are shown as influencing the conditional probabilities, as would be formalized in the structured multivariate measurement models listed above. The following key relationships would be modeled:

- X_{j1} depends on θ_W with a challenge mediated by writing-response demand and global features, Y_{jW} and Y_{j+}, but it does not depend on θ_U. It is possible to construct a good response in terms of language usage alone without really understanding the document. If there is not much of a writing challenge (only a one-word response is required) though, one can't learn much about θ_W.
- X_{j2} depends on θ_U with a challenge mediated by document-processing demand and global features, Y_{jU} and Y_{j+}, and on θ_W to a small extent: It is possible to convey a good understanding of the document with rudimentary language, so some low level of θ_W (to be estimated) is a hurdle to overcome. Once a student has this modicum of writing skill, his value on X_{j2} will depend mainly on θ_U.
- X_{j3} depends on both θ_U and θ_W, with the level of demand for each modeled in terms of the full set of task features. The relationship between θ_U and θ_W for determining X_{j3} would probably be modeled as compensatory, since better understanding and better explanation both contribute to a more effective response.

- X_{j1}, X_{j2}, and X_{j3} are conditionally dependent, all being aspects of a single complex performance. This could be modeled using another task-specific parent variable or modeling dependencies directly, as illustrated in the diagram (Almond & Mislevy, 1999).

[[Figure 2 – document DAG]]

This simple example could be extended in many directions. If more detailed claims about aspects of language use were of interest, then both SM variables θ concerning to those aspects of students' competences, and observable variables X sensitive to the corresponding qualities of the students' responses, would be needed. If different kinds of documents are being addressed and different genres of written response are being required, it would be possible to accumulate evidence about each, based on the appropriate tasks. If this were the case, and if the purpose of the assessment addressed claims concerning competence with different documents and genres, task-model variables indicating *Document Type* and *Response Genre* would be required, and their values in a given task would signal which of the *Document Familiarity* and *Genre Competence* SM variables contribute to performance on that task. If the purpose of the assessment required evidence about the same span of documents and genres but did not need to differentiate students' possibly different proficiencies with different combinations of them, no SM variables would need to be added but students' differing profiles of proficiency for different document types and genres would constitute measurement error about the now more broadly-defined θ_U and θ_W SM variables.

This example sheds some light on the so-called low-generalizability problem often associated with performance assessment (Linn, 1994; Shavelson, et al., 1992). Suppose a set of tasks calls for different mixes of several aspects of knowledge and skill, and students differ in their profiles. If only an overall measure of success is captured for each task, and only an overall tendency to do well is captured as an SM variable, then only the overall level of students' proficiencies accumulates. All of the differences in students' profiles, which may be considerable, is lost in this modeling approach as measurement error. But if the different demands of tasks are modeled, and the differential success of students on different tasks is associated with their different skill profiles, then these

differences can be captured as differing profiles for SM variables. Moreover, reasoning back through the measurement model allows one to project the expected performance of a student with a given profile of SM variable values on a task with a given profile of TM variables. As mentioned above, this is a multivariate generalization of performance-referenced interpretations of IRT variables, and provides support for claims that encompass disparate aspects of competence.

As mentioned in the previous section, models that incorporate effects for raters are an active area in educational measurement. Although generalizability theory (g-theory; Cronbach et al., 1972) has been available since the 1970s, more recent work such as Linacre (1989) FACETS and Patz and Junker (1999) allows us to estimate effects for individual raters. These models are useful not only for taking raters effects into account in inferences about students, but also for monitoring, training, and improving rating in assessment systems (McNamara, 1996). Developments in statistical model-building and estimation[5] make it possible to integrate rater modeling, multivariate student modeling, and structured probabilities depending on task features into a common analysis.

A Second Aspect of Reliability

Earlier we mentioned task scoring as one locus of discussion about reliability. It concerns uncertainty introduced into the argument when going from a work product to the values of observables variables. The second locus is the measurement model, which concerns the amount of information that accumulates over tasks to update the values of student model variables. This issue plays out in terms of reliability in classical test theory, standard errors in IRT, and generalizability indices in g-theory.

Here is a Bayesian perspective on this issue. The probability distribution of observable variables is modeled as a function of the unknown values of SM variables. When the values of observable variables are ascertained, the distributions for SM variables are updated via Bayes Theorem (see, e.g., Bock & Mislevy, 1982, on details for IRT). Starting from a population or an uninformative prior distribution, the posterior distribution for a particular student's SM variables becomes more concentrated as the

evidence from her responses accumulates, as gauged for example by the posterior standard deviation. At any point, it is possible to examine whether there is sufficient accuracy for the purpose of the assessment, be it making important decisions (you need more information for accurate classifications) or guiding instruction (you need less information for low-stakes decisions that can be easily be revised).

Measurement models indicate the direction and the strength of evidence in observations about SM variables, in the form of their mathematical structures and the conditional probability distributions. These principles are by now familiar in IRT, but less so with multivariate models. With multivariate models, it is our experience that working through structures for evidence and task models that have been carefully worked out together by teams of substantive and measurement experts is preferable to creating rich tasks first and hoping that later some one will figure out how to score them and how to model the scores. In TBLA, creating tasks around evidence and task models provides operational efficiencies, allows one to know how to identify and manage evidence, and provides better information about the targeted aspects of student competence.

Task Models

"What tasks or situations should elicit those behaviors?" A task model (TM) provides a framework for constructing and describing the situations in which examinees act. An assessment can have multiple task models, if there are different schemas for these situations. Task model variables describe salient features of tasks. They can play several roles in assessment, including systematizing task construction, focusing the evidentiary value of tasks, guiding assessment assembly, implicitly defining student-model variables, and conditioning the statistical argument between observations and student-model variables (Mislevy, Steinberg, & Almond, in press). A task model includes specifications for the environment in which the student will say, do, or produce something; for example, characteristics of stimulus material, instructions, help, tools, affordances. It also includes specifications for the work product.

---

[5] Specifically, knowledge-based model-construction (Breese et al., 1994) and Markov Chain Monte Carlo estimation (e.g., Gelman et al., 1995).

Writing from the perspective of communicative competence, Skehan (1998, pp. 168-169) says that "if a [task-based assessment] approach is favored, it can only be feasible if we know more about the way tasks themselves influence (and constrain) performance." Initial research has been carried out on features of language-use situations that tend to make them easier or harder, or shift the focus from one aspect of competence to another. Linguistic variables, naturally, were addressed earliest (e.g., Selinker, Tarone, & Hanzeli, 1981), in the context of DSA. Sociolinguistic features that can employed as TM variables concern settings, participants, and purposes (e.g., Bachman, 1990; Bachman & Palmer, 1996; Duran et al., 1985; Skehan, 1998). Features that affect difficulty through cognitive demands include stimulus structure variables and task directives (Mosenthal, 1985; Skehan, 1998; Norris et al., 1998; Robinson, 2001).

We formally specify features as values of task-model variables, defining (possibly infinite, but specifiable) ranges of values, and identifying the (possibly joint) influence of features on targeted aspects of knowledge or capabilities. By identifying and representing key features of TLUs in tasks, assessment designers provide a basis for the claim that tasks are "authentic" to situations outside the assessment itself. By controlling the values of task model variables as features of tasks they create, task authors will have dealt systematically with the focus of tasks' evidence and with the difficulty of the tasks with respect to possibly several interacting aspects of competence. We are far from knowing how all features of tasks interact with all aspects of students' competence. But the best way to leverage what we do know is to build task and evidence models around currently understood relationships.

The Assembly Model

The models described above specify a domain of tasks an examinee might be presented, procedures for evaluating what is observed, and machinery for updating beliefs about the values of the student model variables. *Assembly specifications* define the mix of tasks that will constitute a given student's assessment. One can impose constraints that concern statistical characteristics of tasks, in order to increase measurement precision, or that concern other non-statistical considerations such as

content, format, timing, complexities, cross-item dependencies, and so on (Berger & Veerkamp, 1996).

The Assembly Model is important in our discussion because it manages the interplay among student, task, and evidence models in the following sense: It determines the mix of tasks, through the task model variables (i.e., task features), so that we can (1) determine the range of circumstances that need to be covered to support the targeted claims about the student, thus providing support for validity generalization; (2) control the difficulty of tasks not only overall but with respect to aspects of competence or capabilities in various performance settings; and (3) manage which information tends to accumulate in the form of distributions for SM variables and what information does not accumulate and thus constitutes noise in the statistical model (Almond & Mislevy, 1999).

What Accumulates?

Performance on any task, large or small, entails many aspects of students' knowledge. But simply capturing a work product we believe required some capability does not constitute "measuring" anything. A work product from a given task has the potential to provide evidence about any of those aspects separately or about various amalgamations of them, but it is not until we begin accumulating evidence over items that we can be said to begin the process of measurement.

The idea of accumulating evidence is fundamental to educational measurement (Green, 1978). While every task calls upon a unique mixture of knowledge and skill, the mixtures tapped by two tasks will differ in some ways more than others. Similarities in behaviors in those situations can be attributed to commonalities in the knowledge they demand. Whatever knowledge one item requires that others do not has decreasing influence on synthesized belief as test length increases. In unidimensional models, the accumulation is with respect to a single SM variable. In multidimensional models such as the ones described above, the accumulation is apportioned across multiple SM variables, each in accordance with which SM variables are modeled as being involved in designated aspects of the performance evaluated as observable variables.

A central tenet of ECD is that whatever accumulates over tasks should be intentional rather than accidental. The assessment designer creates the structure of the relationships among tasks, and between tasks and SM variables, through her choices about task features, work products, evaluation rules, observables, measurement models, student-model variables, and test assembly strategies.

The Listening portion of the Test of English as a Foreign Language (TOEFL) is an example. All the items require grasping information from a short talk or conversation in a college environment; each concerns vocabulary and grammar as well as aural processing of language. They don't present clips with background noise, simultaneous speakers, or nonstandard English, so scores don't provide direct evidence about these situations—even though a student would have to deal with these factors in practice. And the items don't require much in the way of cultural knowledge, critical reasoning, or conversational interaction, even though these matters are also central to academic language use. The scores can't say much about these skills, but neither do the considerable differences among students exhibit in these skills add non-accumulating variation, or noise, to scores.

Conclusion

"Data" become "evidence" only when their relevance to some hypothesis, some inference, some claim, is established. In task-based language assessment, this means what we really need to understand first and foremost is the inferential argument associated with the assessment. What is its purpose? What do we want to know, about what students know or can do, in what kinds of situations? Different designers' predilections and perspectives will lead them attack the problem from different starting points. It is the difficulty of fleshing out all the components into a coherent whole that has led to an apparent tension among approaches that go by the names of construct-centered assessment and task-based assessment.

But insights into both constructs and tasks play essential roles in TBLA. A task-centered perspective provides an appropriate starting point for thinking about the features of language-use situations that reveal the language-use competences that are of interest,

and the kinds of performances in those situations that should contain evidence about them. A construct-centered approach will help us think through just what these performances in these situations can tell us about students, in terms *that we must choose* at a level more general than specific performances in specific situations. An evidence-centered perspective guides our construction of student-model elements, measurement models, rubrics, and task-construction frameworks that will make explicit the intuitions that underlie TBLA, and harness them to serve the purpose for which the assessment is intended.

References

Adams, R., Wilson, M.R., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21,* 1-23.

Almond, R.G., & Mislevy, R.J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement,* 23, 223-237.

Almond, R.G., Steinberg, L.S., & Mislevy, R.J. (2001). A sample assessment using the four process framework. CSE Technical Report 543. Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA.

American Council on the Training of Foreign Languages. (1989). *ACTFL proficiency guidelines.* Yonkers, NY: Author.

American Council on the Teaching of Foreign Languages (1999). *ACTFL proficiency guidelines: Speaking (Revised 1999)*. Hastings-on-Hudson: Author.

Bachman, L.F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L.F., & Palmer, A.S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bachman, L.F., & Savignon, S.J. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL Oral Interview. *Modern Language Journal, 70*, 380-389.

Berger, M.P.F., & Veerkamp, W.J.J. (1996). A review of selection methods for optimal test design. In G. Engelhard, & M. Wilson (Eds.), *Objective measurement: Theory into practice (Vol. 3)*. Norwood, NJ: Ablex.

Bock, R.D., & Mislevy, R.J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431-444.

Bradlow, E.T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64,* 153-168.

Breese, J.S., Goldman, R.P., & Wellman, M.P. (1994). Introduction to the special section on knowledge-based construction of probabilistic and decision models. *IEEE Transactions on Systems, Man, and Cybernetics, 24*, 1577-1579.

Brindley, G. (1994). Task-centred assessment in language learning: The promise and the challenge. In N. Bird, P. Falvey, A. Tsui, D. Allison, & A. McNeill (Eds.), *Language and learning: Papers presented at the Annual International Language*

in Education Conference, Hong Kong, 1993 (pp. 73-94). Hong Kong: Hong Kong Education Department.

Brindley, G. (Ed.) (2000). *Studies in immigrant English language assessment.* Sydney: Macquarie University Sydney, National Centre for English Language Teaching and Research.

Brown, J. D., Hudson, T. D., Norris, J. M., & Bonk, W. (in press). *Investigating task-based second language performance assessment.* Honolulu: University of Hawai'i Press.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

DeSmedt, W. (1995). Herr Kommissar: An ICALL conversation simulator for intermediate German. In Holland, V. M., Kaplan, J., & Sams, M. (Eds.), *Intelligent language tutors: Theory shaping technology* (pp. 153-174). Hillsdale, NJ: Lawrence Erlbaum.

Duran,R.P., Canale, M.., Penfield, J., Stansfield, C.S., & Liskin-Gasparro, J.E. (1987). *TOEFL from a communicative viewpoint on language proficiency: A working paper.* TOEFL Research Report No. 17. Princeton, NJ: Educational Testing Service.

Embretson, S.E. (Ed.) (1985). *Test design: Developments in psychology and psychometrics*. Orlando: Academic Press.

Embretson, S. E. (1998). A cognitive design systems approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3*, 380-396.

Fischer, G.H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359-374.

Frase, L.T., Chudorow, M., Almond, R.G., Burstein, J., Kukich, K., Mislevy, R.J., Steinberg, L.S., & Singley, K. (in press). Technology and assessment. In H.F. O'Neil & R. Perez (Eds.), *Technology applications in assessment: A learning view*.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*, 27-32.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. B. (1995). *Bayesian data analysis*. London: Chapman and Hall.

Green, B. (1978). In defense of measurement. *American Psychologist, 33*, 664-670.

Kenyon, D.M., & Malabonga, V. (2001). Comparing examinee attitudes toward computer-assisted and other oral proficiency assessments. *Language Learning and Technology, 5,* 60-83.

Kirsch, I. S., & Jungeblut, A. (1986). *Literacy: Profiles of America's young adults.* Princeton, NJ: National Assessment of Educational Progress/Educational Testing Service.

Linacre, J. M. (1989). *Multi-faceted Rasch measurement.* Chicago: MESA Press.

Linn, R.L. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher, 23*(9), 4-14.

Long, M. H., & Norris, J. M. (2000). Task-based language teaching and assessment. In M. Byram (Ed.), *Encyclopoedia of language teaching* (pp. 597-603).London: Routledge.

McNamara, T. (1996). *Measuring Second Language Performance.* Harlow, Essex, UK: Addison Wesley Longman Ltd.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Education Researcher*, *32*(2), 13-23.

Mislevy, R.J., & Gitomer, D.H. (1996). The role of probability-based inference in an intelligent tutoring system. *User-Modeling and User-Adapted Interaction, 5,* 253-282.

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (in press). On the roles of task model variables in assessment design. In S. Irvine & P. Kyllonen (Eds.), *Generating items for cognitive tests: Theory and practice.* Hillsdale, NJ: Erlbaum.

Mislevy, R.J., Steinberg, L.S., Almond, R.G., Haertel, G., & Penuel, W. (in press). Leverage points for improving educational assessment. In B. Means & G. Haertel (Eds.), *Evaluating the effects of technology in education.* Hillsdale, NJ: Erlbaum.

Mislevy, R.J., Steinberg, L.S., Breyer, F.J., Almond, R.G., & Johnson, L. (in press). Making sense of data from complex assessment. *Applied Measurement in Education.*

Mislevy, R.J., Steinberg, L.S., Breyer, F.J., Almond, R.G., & Johnson, L. (1999). A cognitive task analysis, with implications for designing a simulation-based assessment system. *Computers and Human Behavior, 15,* 335-374.

Mosenthal, P.B. (1985). Defining the expository discourse continuum. *Poetics, 15,*387-414.

Muthén, B. (1988). Some uses of structural equation modeling in validity studies: Extending IRT to external variables. In H. Wainer and H. Braun (Eds.), *Test validity.* Hillsdale, NJ: Erlbaum.

Norris, J. M., Brown, J. D., Hudson, T. D., & Yoshioka, J. K. (1998). *Designing second language performance assessment.* Honolulu: University of Hawai'i Press.

Patz, R. J., & Junker, B. W. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, *24*(4), 342-366.

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics, 22*, 27-57.

Selinker, L., Tarone, E., & Hanzeli,V. (Eds.) (1981). *English for technical and academic purposes: Studies in honor of Louis Trimble.* Rowley, MA: Newbury House.

Shavelson, R.J., Baxter, G.P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher, 21*(4), 22-27.

Sheingold, K. , & Frederiksen, J. R. (1994). Using technology to support innovative assessment. In B. Means (Ed.), Technology and education reform (pp. 111-131). San Francisco: Jossey-Bass.

Skehan, P. (1998). *A cognitive approach to language learning.* Oxford: Oxford University Press.

Wiggins, G. (1993). *Assessing student performance.* San Francisco: Jossey Bass.

Wolf, D., Bixby, J., Glenn, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. In G. Grant (Ed.), *Review of Educational Research, Vol. 17* (pp. 31-74). Washington, DC: American Educational Research Association.

Table 1

Excerpts from the ACTFL Proficiency Guidelines for Reading*

Level	Generic Description
Novice-Low	Able occasionally to identify isolated words and/or major phrases when strongly supported by context.
:	:
Intermediate-Mid	Able to read consistently with increased understanding simple connected texts dealing with a variety of basic and social needs.… They impart basic information about which the reader has to make minimal suppositions and to which the reader brings personal information and/or knowledge. Examples may include short, straightforward descriptions of persons, places, and things, written for a wide audience.
:	:
Advanced	Able to read somewhat longer prose of several paragraphs in length, particularly if presented with a clear underlying structure. … Comprehension derives not only from situational and subject matter knowledge but from increasing control of the language. Texts at this level include descriptions and narrations such as simple short stories, news items, bibliographical information, social notices, personal correspondence, routinized business letters, and simple technical material written for the general reader.
Advanced-Plus	…Able to understand parts of texts which are conceptually abstract and linguistically complex, and/or texts which treat unfamiliar topics and situations, as well as some texts which involve aspects of target-language culture. Able to comprehend the facts to make appropriate inferences. …
Superior	Able to read with almost complete comprehension and at normal speed expository prose on unfamiliar subjects and a variety of literary texts. Reading ability is not dependent on subject matter knowledge, although the reader is not expected to comprehend thoroughly texts which are highly dependent on the knowledge of the target culture.… At the superior level the reader can match strategies, top-down or bottom-up, which are most appropriate to the text.…

* Excerpts from the *ACTFL proficiency guidelines,* American Council on the Training of Foreign Languages (1989).

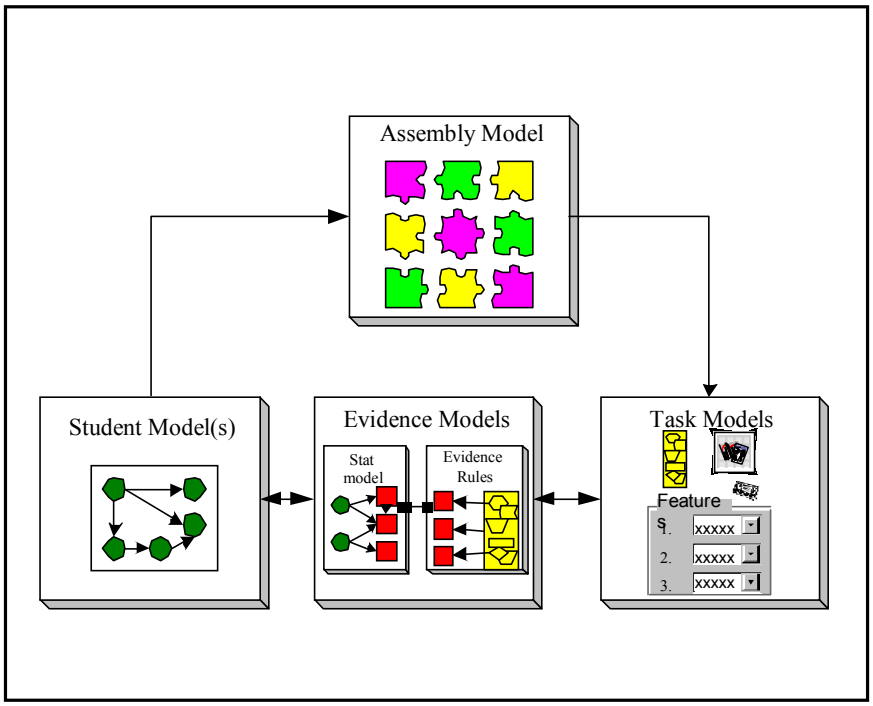

Figure 1
Main Models of the Conceptual Assessment Framework

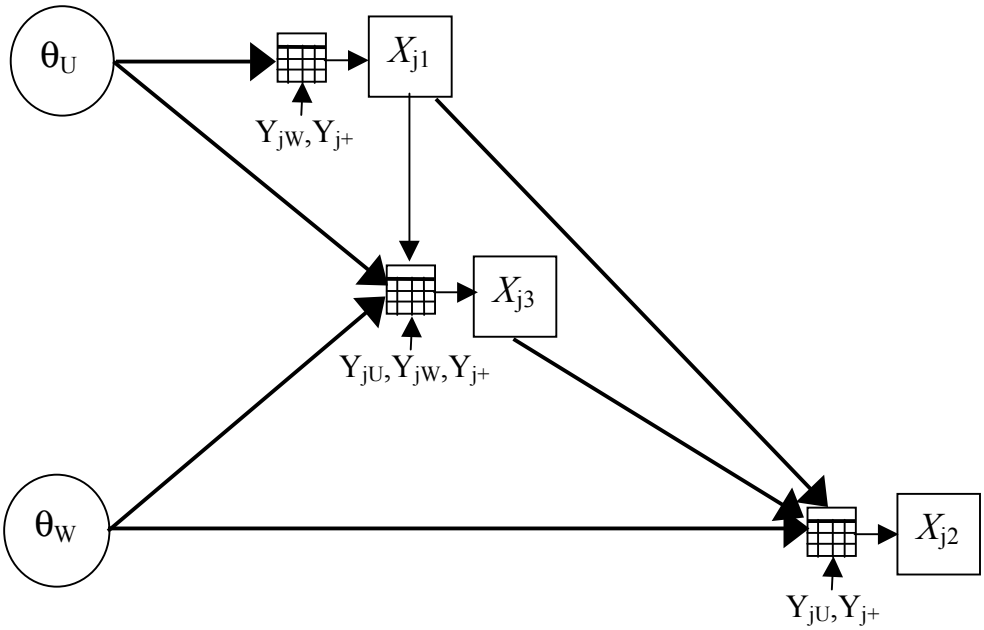

Figure 2

Graphical Representation of the Measurement Model for a Task with Three Observable Variables