Design and analysis in task-based language assessment

Robert J. Mislevy *University of Maryland,* **Linda S. Steinberg** and **Russell G. Almond** *Educational Testing Service*

In task-based language assessment (TBLA) language use is observed in settings that are more realistic and complex than in discrete skills assessments, and which 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 article examines these challenges from the perspective of 'evidence-centred assessment design'. The main building blocks 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.

I Introduction

Task-Based Language Assessment (TBLA)[1] is '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: 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 and 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 broadens consideration to the social

Address for correspondence: Robert J. Mislevy, University of Maryland, 1230-C Benjamin Building, College Park, MD 20742, USA; email: rm257@umail.umd.edu

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

Language Testing 2002 19 (4) 477–496 10.1191/0265532202lt241oa

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 second language competence (e.g., Bachman and Palmer, 1996), 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.

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

II 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 does in the assessment to what the student knows or can do, but with direct evidence about only very focused and limited uses of language. The problem is that a purpose for assessing may require 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 second language 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 is not enough simply to 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 is not enough either. Developing psychometric models that deal with more complex data may be necessary too, but it still is not 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, of course, practical benefits to explicating an assessment argument. 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, on the one hand, and connections with target language use situations, on the other (Messick, 1994).

Generally applicable principles of evidentiary reasoning can 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 what we are learning about the ways that language knowledge interacts with other knowledge to constitute ability for use (McNamara, 1996: 48). Evidence-centred design offers a framework for first working through the structure of the argument, then designing elements that can be assembled to transform that argument into an operational assessment.

We would argue that these evidentiary-reasoning principles must underlie any coherent assessment, and are implicit in common practices that do produce assessments that suit their purposes in familiar ways. There is much to be gained by working from an explicit normative model such as ECD, however, when designing assessments for which off-the-shelf design practice simply will not suffice: e.g., assessments using richer tasks, tackling more ambitious purposes, gathering new kinds of data or requiring more complex measurement

models. In informal uses of TBLA, such as in the language-learning classroom, the salient aspect of ECD is the framework for working through the rationale for creating tasks and evaluating performance, not measurement models. Simple measurement models can suffice when an assessment is embedded in an instructional context. However, for large scale, formal assessments that are characterized by high stakes and large samples, special attention must be accorded to coordinating language-learning theory and measurement models. In these applications, the interplay among substantive considerations, statistical models and patterns in examinees' data leads iteratively to improved practice.

III The ECD framework

This section briefly introduces the basic high-level models of the evidence-centered design framework. In a short article such as this we can offer neither a complete presentation nor a 'how-to-do-it' guide,[2] but we can describe the high-level structure that connects an assessment argument to the nuts and bolts of operational assessment. Figure 1 is a high-level schematic of four basic models in a 'conceptual assessment framework': Student models,[3] evidence models and task models, along with an assembly model that governs the way that individual tasks are brought together to form assessments. In brief, the student model specifies the variables in terms of which we wish to characterize students. Task models are schemas for ways to elicit 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 students' performances in individual task situations – i.e., ascertaining the values of observable variables – and the measurement component contains the statistical machinery for updating beliefs about student-model variables accordingly. An operational assessment will generally have one student model, but can use multiple task and evidence models to produce data that provide evidence about different aspects of skill and knowledge. An assembly model specifies how individual tasks are combined to produce an assessment.

---

[2]Unfortunately, this is especially true with respect to applications requiring multivariate measurement models, which we feel are sorely needed for complex mixtures of TBLA tasks. We strongly urge the interested reader to follow up on references provided in this connection.

[3]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.

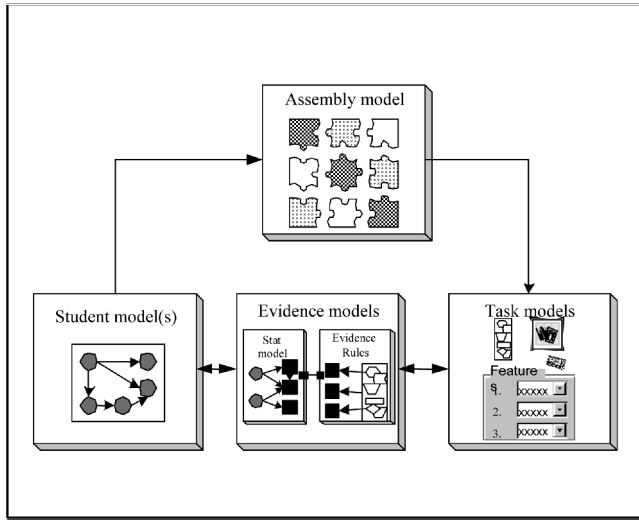

Figure 1 Main models of the conceptual assessment framework

IV The student model

What complex of students' knowledge, skills or other attributes should be assessed? The student model (SM) in Figure 1 depicts student model variables as circles. Configurations of their values are meant to approximate – as appropriate to the purpose of the assessment – 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. We cannot see the values of SM variables directly. We see what students say or do, and construe that as evidence about SM variables. 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 the light of new evidence (Mislevy and Gitomer, 1996). Psychometric models such as classical test theory (CTT), item response theory (IRT), factor analysis and latent class models are consonant with this approach.

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 use the term here, refers to a piece of machinery: a set of variables in a probability model, for accumulating evidence about students. Operationally, SM variables summarize patterns of values of observable variables, along lines we build into the evidence-model structures (discussed below), as evoked by features we build into tasks (also discussed below).

By way of contrast, a model such as that in Bachman and Palmer

(1996) 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 do not determine the statistical student model that is appropriate for a given assessment. The purpose of the assessment is further required to determine the focus and the grain size of the variables in the statistical student model.

1 The relationship between claims and student-model variables

Claims in assessment are substantively meaningful statements we would like to make about what students know, can do or have accomplished, explicitly or implicitly in relation to contexts, as required to suit the purpose of the assessment. For example, consider a language assessment for medical students for whom English is a second language (McNamara, 1996). Treating and counselling English-speaking patients is the targeted language use (TLU; Bachman and Palmer, 1996), and having students work through consultation scenarios with English speaking simulated patients provides direct evidence about their capabilities to do this, in a 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 at her clinic. A school advisor would need finer-grained claims, because just knowing that a student could not handle consultation in English does not indicate whether the student's difficulties lie with language use *per se*, or with medical knowledge or interpersonal skills. His or her claims would need to be at a level that would guide placement into practice sessions.

What is the relationship between substantive claims and student-model variables? Here are four ways an assessment designer can establish the relationship between them.

One approach posits a 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. 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%.'

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). The ACTFL guidelines move from the level 'novice low', in which the student can typically only use language in a given modality in rudimentary ways, up to 'superior'. Each level is described in terms of several

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 . . .

Source: ACTFL, 1989

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.

A third alternative for addressing multiple claims with a single SM variable is to model response probabilities for tasks with particular key features. An example is the Document Literacy scale from the Young Adult Literacy Survey (YALS; Kirsch and Jungeblut, 1986), in which:

- tasks requiring simple responses are written in accordance with numbers or degrees of features that are salient under the YALS cognitive model, such as the kind and structure of the documents and the complexity of what the student is asked to do;
- a single IRT variable θ accounts well for performance across all the tasks; and
- a regression model using salient features as predictors accounts well for task difficulties.

Substantive meaning is imparted to any given level of θ by describing the chances of success of students at that level in situations with various features.

A fourth approach tackles the interactions among competences and contexts, as 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 profiles 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 tasks that stress competences in different ways. 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 given match-up can be described in terms of values of observable variables.

For an example of this multivariate approach, we use the Document Literacy scale discussed above and introduce an additional scale concerning 'written explanation'. We saw above how 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 both provide documents of varying complexities and require written responses of varying complexities. For example, we can have simple documents and ask for simple phrasal 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 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. 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 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.

V Evidence models

What observable behaviours or performances might provide evidence about the knowledge or skills we wish to measure? Evidence models lay out the argument about why and how observations in a given task situation constitute evidence about student model variables. There are two parts to an evidence model, the evaluation component and the measurement component.

1 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; e.g., 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. The observable variables are evaluative summaries of what the assessment designer has determined to be the key aspects of the performance. Evaluation rules indicate how to carry out these mappings. 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 can be evaluated, and multiple evaluations of either or both can be made.

As an example, consider a learner's oral directions to the library. We might rate its overall effectiveness, or we could further distinguish the aptness of its content, its complexity, the accuracy of its grammar and vocabulary and its sociolinguistic appropriateness. These latter qualities tend to trade off against one another in use, as

people decide how to get the most benefit from the competences they possess (Skehan, 1998: 167ff).[4] What we should capture with evaluation rules, therefore, depends on which aspects of competence, usage and effectiveness are needed to serve the assessment's purpose. The overall evaluation could suffice for predicting effectiveness in a job setting, but feedback on particular competences would require ratings that captured evidence about aspects of the performance that focus on those competences.

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 behaviours we need to see, in situations with what kinds of features, we can consider a range of possible ways to make and evaluate observations. 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 gain direct evidence about a student's capability to do this in given situations, we need to observe him or her reacting to a conversation partner, and adding to the interaction in a way that adds 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 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). The trade-offs among evidentiary value, costs and consequences must be examined in light of an assessment's purposes and constraints.

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 and 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 is the incorporation of rater effects into measurement models. While generalizability theory (g-theory; Cronbach *et al.*, 1972) has been available since the 1970s,

[4]Knowing about trade-offs argues for modelling these multiple ratings as conditionally dependent (Bradlow *et al.*, 1999).

more recent work such as Linacre's (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).

2 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, with its single SM variable θ that gauges an overall proficiency in the set of tasks, and factor analysis, with its multivariate student model discovered from patterns in the data.

An alternative approach of particular relevance to TBLA uses a multivariate student model and tasks designed to elicit evidence about particular SM variables through choices of task features. 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 to mediate the stress put on those aspects. These relationships can be built into statistical models to make explicit the expected relationships, guide task construction and exploit the structure in inferences about SM variables (Embretson, 1985). Suitable measurement models include item-level confirmatory factor analysis (e.g., Muthén, 1988) structured IRT models (e.g., Adams *et al*., 1997; Embretson, 1998), and modular assemblies of Bayes net fragments (e.g., Almond and Mislevy, 1999; Mislevy, Steinberg, Breyer, Almond and Johnson, in press).

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., Norris *et al*., 1998; Skehan, 1998; Brindley, 2000; Brown *et al*., in press). 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. As in any statistical modelling effort, we monitor the fit of the model for concordance among the patterns in the data, the measurement model and our notions about the nature of competence and performance in the situations we have designed.

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 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 denote these as Y_{jW}. Other features of the task affect performance globally, such as time limit; we 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 can be modelled 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 modelled:

- 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), one cannot 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

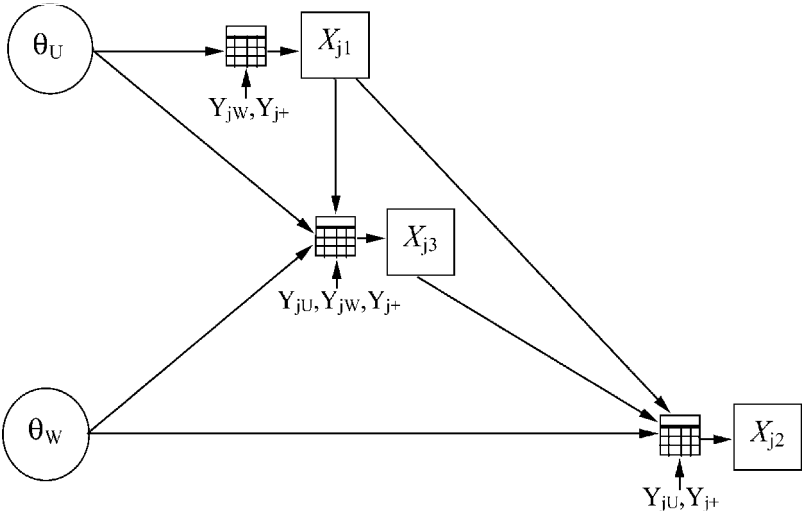

Figure 2 Graphical representation of the measurement model for a task with three observable variables

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 or her 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 modelled in terms of the full set of task features. The relationship between θ_U and θ_W for determining X_{j3} would probably be modelled 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 modelled using another task-specific parent variable or modelling dependencies directly, as illustrated in the diagram (Almond and Mislevy, 1999).

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, then 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 (Shavelson *et al*., 1992; Linn, 1994). 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 is lost in this modelling approach as measurement error. But if the different demands of tasks are modelled, 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.

A second aspect of reliability: Earlier we mentioned task scoring as one locus of discussion about reliability. It concerns uncertainty introduced when we reason from a work product to the values of observable variables. The second locus of reliability is the measurement model, which concerns the amount of information that accumulates over tasks to update beliefs about student model variables. From a Bayesian perspective, the probability distribution of observable variables is modelled 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. Starting from an uninformative prior distribution, the posterior distribution for a student's SM variables becomes more concentrated as the evidence from his or her responses accumulates, as gauged for example by the posterior standard deviation. At any point, we can see 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).

VI Task models

What tasks or situations should elicit the behaviours we need as evidence? A task model (TM) is a schema 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 and 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.

Writing from the perspective of communicative competence, Skehan (1998: 168–69) says that 'if a [task-based assessment] approach is favoured, 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 were, naturally, addressed earliest (e.g., Selinker *et al*., 1981), in the context of DSA. Sociolinguistic features that can be employed as TM variables concern settings, participants and purposes (e.g., Duran *et al*., 1985; Bachman, 1990; Bachman and Palmer, 1996; Skehan, 1998). Features that affect difficulty through cognitive demands include stimulus structure variables and task directives (Mosenthal, 1985; Norris *et al*., 1998; Skehan, 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.

VII 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 constitute a given student's assessment. One can impose constraints that concern statistical characteristics of tasks, to increase measurement precision, or that concern nonstatistical considerations such as content, format, timing, complexities, cross-item dependencies and so on (Berger and Veerkamp, 1996).

The Assembly Model manages the interplay among student, task and evidence models: It determines the mix of tasks, through the task model variables (i.e., task features), so we can:

- determine the range of circumstances that need to be covered to support the targeted claims about the student, and thus support validity generalization;
- control the difficulty of tasks both overall and with respect to various aspects of competence or capabilities; and
- manage what 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 and Mislevy, 1999).

Performance on any task, large or small, entails many aspects of students' knowledge. A work product from a given task has the potential to provide evidence about any of those aspects, but it is not until we begin accumulating evidence over items that we can be said to begin the process of measurement (Green, 1978). While every task calls upon a unique mixture of knowledge and skill, similarities in examinees' behaviours across certain tasks can be attributed to commonalities in the knowledge they demand. Whatever knowledge one item requires that others do not has decreasing influence on accumulating beliefs as test length increases. In unidimensional models, the accumulation is with respect to a single SM variable. In multidimensional models, the accumulation is apportioned across multiple SM variables, according to which SM variables are modelled as being producing aspects of the performance captured in observable variables.

VIII Conclusions

'Data' become 'evidence' only when their relevance to some hypothesis, some inference, some claim, is established. In task-based

language assessment, this means that 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 will lead them to attack the problem from different starting points. The difficulty of fleshing out all the components into a coherent whole has led to an apparent tension among design 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 is 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 helps us think through just what these performances in these situations can tell us about students, at a level above specific performances in specific situations. An evidence-centered perspective guides our construction of student-model elements, measurement models, rubrics and task-construction frameworks that make explicit the intuitions that underlie TBLA, and harnesses them to serve the purpose for which the assessment is intended.

Acknowledgements

We are grateful to John Norris and Lyle Bachman for comments on earlier versions of this article, 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, US 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 US Department of Education.

IX References

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

—— 1999: *ACTFL proficiency guidelines: speaking*. Revised 1999. Hastings-on-Hudson: Author.

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

Almond, R.G. and **Mislevy, R.J.** 1999: Graphical models and computerized adaptive testing. *Applied Psychological Measurement* 23, 223–37.

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

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

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

Berger, M.P.F. and **Veerkamp, W.J.J.** 1996: A review of selection methods for optimal test design. In Engelhard, G. and Wilson, M., editors, *Objective measurement: theory into practice, Volume 3*. Norwood, NJ: Ablex.

Bradlow, E.T., Wainer, H. and **Wang, X.** 1999: A Bayesian random effects model for testlets. *Psychometrika* 64, 153–68.

Brindley, G. 1994: Task-centred assessment in language learning: the promise and the challenge. In Bird, N., Falvey, P., Tsui, A., Allison, D. and McNeill, A., editors, *Language and learning: papers presented at the Annual International Language in Education Conference (Hong Kong, 1993)*. Hong Kong: Hong Kong Education Department, 73–94.

—— editor, 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. and **Bonk, W.** in press: *Investigating task-based second language performance assessment.* Honolulu: University of Hawaii Press.

Cronbach, L.J., Gleser, G.C., Nanda, H. and **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. and Sams, M., editors, *Intelligent language tutors: theory shaping technology.* Hillsdale, NJ: Lawrence Erlbaum, 153–74.

Duran, R.P., Canale, M., Penfield, J., Stansfield, C.S. and **Liskin-Gasparro, J.E.** 1985: *TOEFL from a communicative viewpoint on language proficiency: a working paper.* TOEFL Research Report No. 17. Princeton, NJ: Educational Testing Service.

Embretson, S.E., editor, 1985: *Test design: developments in psychology and psychometrics*. Orlando, FL: Academic Press.

—— 1998: A cognitive design systems approach to generating valid tests: application to abstract reasoning. *Psychological Methods* 3, 380–96.

Fischer, G.H. 1973: The linear logistic test model as an instrument in educational research. *Acta Psychologica* 37, 359–74.

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

Frederiksen, J. R. and **Collins, A.** 1989: A systems approach to educational testing. *Educational Researcher* 18, 27–32.

Green, B. 1978: In defense of measurement. *American Psychologist* 33, 664–70.

Kirsch, I.S. and **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, IL: MESA Press.

Linn, R.L. 1994: Performance assessment: policy promises and technical measurement standards. *Educational Researcher* 23, 4–14.

Long, M.H. and **Norris, J.M.** 2000: Task-based language teaching and assessment. In Byram, M., editor, *Encyclopedia of language teaching*. London: Routledge, 597–603.

McNamara, T. 1996: *Measuring second language performance*. Harlow:

Messick, S. 1994: The interplay of evidence and consequences in the validation of performance assessments. *Education Researcher* 32, 13–23.

Mislevy, R.J. and **Gitomer, D.H.** 1996: The role of probability-based inference in an intelligent tutoring system. *User-Modeling and User-Adapted Interaction* 5, 253–82.

Mislevy, R.J., Steinberg, L.S., Breyer, F.J., Almond, R.G. and **Johnson, L.** 1999: A cognitive task analysis, with implications for designing a simulation-based assessment system. *Computers and Human Behavior* 15, 335–74.

Mislevy, R.J., Steinberg, L.S. and **Almond, R.G.** in press: On the structure of educational assessments. *Measurement*.

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

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 Wainer, H. and Braun, H., editors, *Test validity.* Hillsdale, NJ: Erlbaum.

Norris, J.M., Brown, J.D., Hudson, T.D. and **Yoshioka, J.K.** 1998: *Designing second language performance assessment.* Honolulu: University of Hawaii Press.

Patz, R.J. and **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, 342–66.

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. and **Hanzeli, V.**, editors, 1981: *English for*

technical and academic purposes: studies in honor of Louis Trimble. Rowley, MA: Newbury House.

Shavelson, R.J., Baxter, G.P. and **Pine, J.** 1992: Performance assessments: political rhetoric and measurement reality. *Educational Researcher* 21, 22–27.

Sheingold, K. and **Frederiksen, J.R.** 1994: Using technology to support innovative assessment. In Means, B., editor, *Technology and education reform.* San Francisco, CA: Jossey-Bass, 111–31.

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

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

Wolf, D., Bixby, J., Glenn, J. and **Gardner, H.** 1991: To use their minds well: investigating new forms of student assessment. In Grant, G., editor, *Review of educational research, Volume 17.* Washington, DC: American Educational Research Association 31–74.