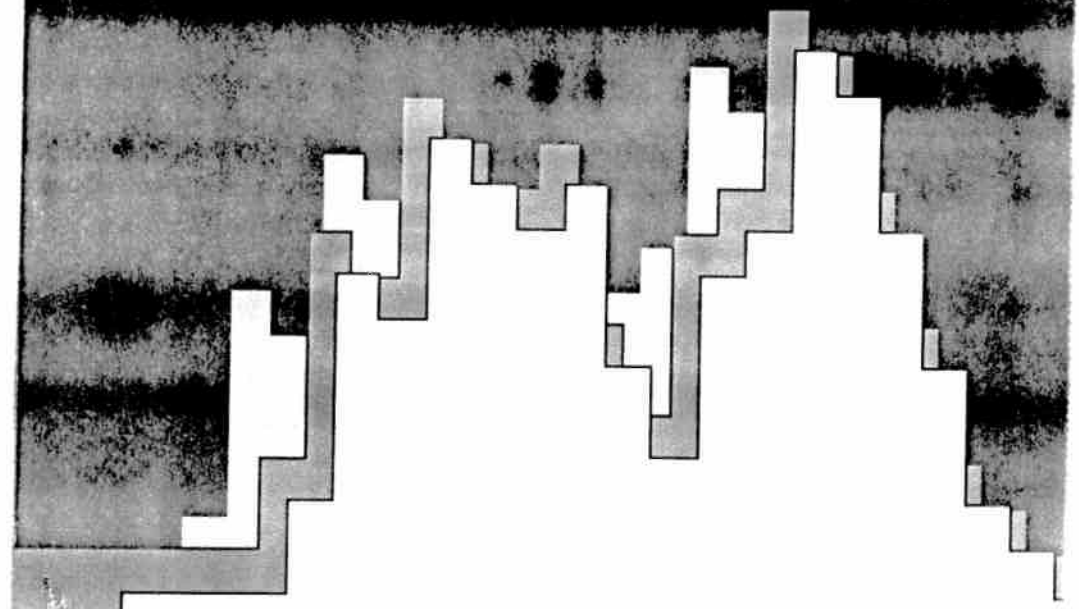
Mueller, R.J. (1995). Test theory and language learning assessment.

LANGUAGE TESTING

EDITORS:
Alan Davies and John Upshur



Volume 12 Number 3 1995
ISSN 0265-5322

EDWARD ARNOLD

Test theory and language-learning assessment *

Robert J. Mislevy *Educational Testing Service*

Standard test theory is machinery for carrying out inference in a particular admixture of ideas from statistics, measurement and psychology that coalesced in the first third of this century. Recent developments in cognitive and educational psychology, such as increased appreciation of the situated nature of learning and understanding, call for broader ranges of student models and types of data. Just as under the standard testing paradigm, however, we face such questions as: what kinds of evidence are needed to support inferences about students? How much faith can we place in the evidence, and in the statements? Are elements of evidence overlapping, redundant or contradictory? When must we ask different questions or pose additional situations to distinguish among competing explanations of what we see? A conceptualization of test theory is discussed which is meant to address issues of weight and coverage of evidence for statements framed in more recent educational/psychological paradigms. Implications for language assessments built around the ACTFL guidelines are considered.

HOLMES: In solving a problem of this sort, the grand thing is to be able to reason backward. That is a very useful accomplishment but people do not use it much. In everyday affairs of life it is more useful to reason forward, and so the other comes to be neglected. There are fifty who can reason synthetically for one who can reason analytically.
WATSON: I confess I do not follow you.
HOLMES: I hardly expected that you would. Let me see if I can make it clearer. Most people, if you describe a train of events to the will tell you what the results would be. They can put those events together in their mind, and argue from them that something will come to pass. There are few, however, who, if you told them a result, would be able to evolve from their own inner consciousness what the steps were which led up to that result. This power is what I mean when I talk about reasoning backward, or analytically.
WATSON: I understand (Doyle, 1930: 268).

I Introduction

Test theory, as we usually think of it, is part of a package. It encompasses models and methods for drawing inferences about

* Plenary address at the Center for the Advancement of Language Learning's 1994 Language Aptitude Invitational Symposium, 25–27 September, Arlington, VA. I am grateful for discussions with Nancy Anderson, Dan Eignor, Anne Harvey and Ming Mae Wang.

what students know and can do – as cast in a particular framework of ideas from measurement, education and psychology that coalesced in the first third of the twentieth century. In a nutshell, 1) human abilities were viewed as traits, or '... relatively stable characteristics of a person – attributes, enduring processes, or dispositions – which are consistently manifested to some degree when relevant, despite considerable variation in the range of settings and circumstances' (Messick, 1989: 15); 2) traits were conceived as numbers along measurement scales, locating people along continua of mental characteristics just as their heights and weights located them along continua of physical characteristics; 3) tendencies in behaviour in samples of a domain of discrete settings and circumstances (e.g., assessment tasks) were the privileged form of evidence about traits; and 4) the purpose of test theory was to guide reasoning from observed behaviour in samples of situations from the domain to inferences about traits.

This 'domain-behaviour' framework of assessment generates a universe of discourse: the nature of the problems one perceives, the kinds of statements one makes about students, the ways one gathers data to support them. Test theory, as we usually think of it, is the application of inferential principles to deal with such problems as missing data, source unreliability, multistage inference, conflicting or overlapping observations, multiple sources of disparate evidence, and constrained resources for gathering and evaluating information – as they arise in this framework.

The views of the nature and the acquisition of competence in a second language, and the nature of inferences we would wish to make about students' developing competence, do not always fall within this familiar realm. In particular, we may wish to take into account the situated and contextual aspects of language learning, and we may wish to gather data from complex tasks that stress the interconnections among aspects of students' competence. But while these developments may suggest student models and observational strategies quite different from those employed by Spearman, Thurstone and Thorndike, practical work under alternative perspectives inevitably faces in some form the same general inferential problems listed above. This is where a more broadly construed conception of test theory is required. It is not sufficient merely to define the class of conjectures about student competence we wish to address, and devise settings in which students can display these competencies. We must, further, specify how what we observe is related to competence as we choose to conceive it, and construct a framework for carrying out inference within the framework we thus erect.

To this end, the following section discusses the notions of

evidence and inference more broadly than they are usually conceived in educational assessment. The role of probability-based inference in assessment is described. Ideas are then illustrated with two language-learning assessment challenges – contextual effects on learning and complex performance tasks – with regard to inference in the conceptual framework of the American Council on the Training of Foreign Languages (ACTFL) guidelines (ACTFL, 1989).

II Evidence and inference

Inference is reasoning from what we know and what we observe to explanations, conclusions or predictions. The skills we must apply in educational assessment are essentially the same as those employed in such fields as troubleshooting, medical diagnosis, criminology and intelligence analysis. We attempt to establish the weight and coverage of evidence in what we observe. The very first question we must address is 'Evidence about what?' Schum (1987: 16) points out the crucial distinction between *data* and *evidence*: 'A datum becomes evidence in some analytic problem when its *relevance* to one or more hypotheses being considered is established . . . [E]vidence is relevant on some hypothesis if it either increases or decreases the likeliness of the hypothesis. Without hypotheses, the relevance of no datum could be established.'

Test data acquire meaning only in relation to particular hypotheses, or conjectures, that we entertain. The same observation can be direct evidence for some conjectures and indirect evidence for others, and wholly irrelevant to still others. In educational assessment, we construct our conjectures around notions about the nature and the acquisition of competence. We can actually observe only the specific actions and products that students produce in specific circumstances. To evaluate their progress or guide further instruction, however, we talk at a higher level of abstraction, using specific observations as evidence for our inferences.

A conception of competence is effected as a set of variables in a student model, a simplified description of selected aspects of the infinite varieties of skills and knowledge that characterize real students. Depending on our purposes, we might distinguish anywhere from one or hundreds of facets. They might be expressed in terms of numbers, categories or some mixture; they might be conceived as persisting over long periods of time, or apt to change at the next problem-step. They might concern tendencies in behaviour, conceptions of phenomena, available strategies or levels of development. The point is that we don't observe these variables directly. We

observe only student's behaviour in limited circumstances – indirect evidence about competence more abstractly conceived. Test theory, broadly construed, is conceptual and statistical machinery for reasoning from observations to inferences in terms of the competence model.

Suppose we want to make a statement about Jasmine's proficiency, in terms of likely values of the variables in a model built around some key aspects of competence. We can't observe these values directly,[1] but perhaps we can make an observation that bears information about the plausibility of various values under the model: her answer to a multiple-choice question, say, or two sets of judges' ratings of her violin solo, or an essay outlining how to determine which paper towel is most absorbent. The observation can't tell us her value with certainty, because similar behaviour could be produced by students with different underlying levels of competency depending on factors such as their familiarity with the context and situation. It is, however, more likely to be produced by students at some levels than others. Nonsensically answering '¿Como está usted?' with 'Me llamo Carlos', for example, is much more likely from a student classified as a 'low-novice' under the ACTFL reading guidelines (see Table 1) than an 'advanced' student. (There are similarly conceived guidelines for writing, speaking and listening.)

III Conceptions of competence

A conception of student competence and a purpose for assessment should drive the particular methods we need to get students to act in ways that reveal something about their competencies, or the forms of assessment we employ. This section contrasts key aspects of two broadly cast assessment paradigms, which we shall refer to as the 'domain-behaviour' and 'cognitive/developmental' paradigms, and notes some implications for assessment forms and test theory.

The 'domain-behaviour' paradigm originated under trait psychology and evolved further under behaviourist psychology. From trait psychology came the notions of characterizing characteristics of persons in terms of numbers on a measurement scale, and taking as evidence for these numbers counts of keyed behaviours in samples from a domain of relevant settings (such as test items). The following quotation reflects how this perspective came to be applied

[1] After all, the model itself isn't truth but a simplified approximation we have constructed, and variable values are not so much characteristics of Jasmine but of summaries of our knowledge about patterns we perceive in Jasmine's behaviour, as seen through the lens of the model.

to the development and practice of educational assessment:

> The educational process consists of providing a series of environments that permit the student to learn new behaviors or modify or eliminate existing behaviors and to practice these behaviors to the point that he displays them at some reasonably satisfactory level of competence and regularity under appropriate circumstances. The statement of objectives becomes the description of behaviors that the student is expected to display with some regularity. The evaluation of the success of instruction and of the student's learning becomes a matter of placing the student in a sample of situations in which the different learned behaviors may appropriately occur and noting the frequency and accuracy with which they do occur (Krathwohl and Payne, 1971: 17–18).

Under the domain-behaviour approach, the specification of an

Table 1 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 [emphasis added]'
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 [emphasis added]'
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 [emphasis added]'
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: Based on ACTFL (1989).

assessment describes a collection of task contexts as seen from the assessor's point of view, and provides a system for classifying the responses students might make. Potential responses in some contexts, such as multiple-choice items, are unambiguously right or wrong; in others, counts or instances of behaviours of certain types, the distinction of which may require expert judgement, are recorded. Behaviour observed in a sample of tasks constitutes direct evidence for expected behaviour in the domain as a whole, which in turn constitutes an operational definition of competence. The primary inferential task of standard test theory is to characterize the weight of evidence that samples of tasks provide about students' domain proficiencies. The processes by which students *acquire* competence are of interest, of course, to students, teachers and researchers alike, but for the most part these questions lie outside the universe of discourse associated with the domain-behaviour paradigm of assessment (Stake, 1991).

In contrast, the acquisition of competence plays a central role in contemporary cognitive and educational psychology. The following quotation reflects the cognitive/developmental perspective as it relates to educational assessment:

> Essential characteristics of proficient performance have been described in various domains and provide useful indices for assessment. We know that, at specific stages of learning, there exist different integrations of knowledge, different forms of skill, differences in access to knowledge, and differences in the efficiency of performance. These stages can define criteria for test design. We can now propose a set of candidate dimensions along which subject-matter competence can be assessed. As competence in a subject-matter grows, evidence of a knowledge base that is increasingly *coherent*, *principled*, *useful*, and *goal-oriented* is displayed, and test items can be designed to capture such evidence (Glaser, 1991: 26, emphasis in original).

From the cognitive perspective, the specifications for an assessment describe contexts that can evoke evidence about students' competence as conceived at a higher level of abstraction, and provide judgemental guidelines for mapping from observed behaviour to this inferred competence. This behaviour provides evidence about competence so conceived, but not necessarily *direct* evidence. We may have to interpret this behaviour in the light of additional knowledge or supporting evidence about, for example, how the content or the context of a task interacts with the student; we may need to infer, or learn more about, the task as seen from the point of view of the student.

The ACTFL reading proficiency guidelines (Table 1) illustrate this point. Contrast the description of intermediate readers' competence with texts '. . . about which the reader has personal interest or

knowledge' with advanced readers' competence with '... texts which treat unfamiliar topics and situations'. This distinction is fundamental to the underlying conception of developing language proficiency, but obviously a situation that is familiar to one student is unfamiliar to others. The evidential import of the same behaviour in the same situation can differ radically for different students and, as we shall explore further, affect what we infer about their capabilities from their behaviour.

IV Probability-based inference

> Probability isn't really about numbers; it's about the structure of reasoning (Glenn Shafer, quoted in Pearl, 1988: 44).

As the preceding section addressed what we want to reason about in educational assessment, this section concerns how we want to reason. It outlines the basic kinds of reasoning tasks we face, and reviews some tools from probability theory we can gainfully employ to this end, some hundreds of years old and others quite recent.

1 Kinds of inference

Schum (1987) distinguishes among deductive, inductive and abductive reasoning, all of which play essential and interlocking roles in educational assessment:

- *Deductive reasoning* flows from generals to particulars, within an established framework of relationships among variables – from causes to effects, from diseases to symptoms, from the way a crime is committed to the evidence likely to be found at the scene, from a student's knowledge and skills to observable behaviour. Under a given state of affairs, what are the likely outcomes?
- *Inductive reasoning* flows in the opposite direction, also within an established framework of relationships – from effects to possible causes, from symptoms to possible diseases, from a student's solution to likely configurations of knowledge and skill. Given the outcomes we see, what state of affairs may have produced them?
- *Abductive reasoning* (a term coined by the philosopher Charles C. Peirce) proceeds from observations to new hypotheses, new variables or new relationships among variables. 'Such a "bottom-up" process certainly appears similar to induction; but there is an argument that such reasoning is, in fact, different from induction since an existing hypothesis collection is enlarged in the process. Relevant evidentiary tests of this new hypothesis

are then *deductively* inferred from the new hypothesis' (Schum, 1987: 20).

Conjectures, and the understanding of what constitutes evidence about them, emanate from the variables, concepts and relationships of the field within which reasoning is taking place. The theories and explanations of a field suggest the structure through which deductive reasoning flows – the 'generative principles of the domain', to borrow a phrase from Greeno (1989). Inductive and abductive reasoning depend just as critically on the same structures, as the task is to speculate on circumstances which, when their consequences are projected deductively, lead plausibly to the evidence at hand. Determining promising possibilities, we reason deductively to other likely consequences – potential sources of corroborating or disconfirming evidence for our conjectures.

A detective at the scene of a crime reasons abductively to reconstruct the essentials and principals of the event. Anything he sees, in the light of a career of experience, can suggest possibilities; ways things might have happened which, reasoning deductively, could have produced the present state of affairs (e.g., documents, testimony, physical evidence). Given tentative hypotheses, does inductive reasoning from other observations conflict or fit in? When they conflict, does their juxtaposition spark a new hypothesis? A successful investigation leads to a plausible explanation of the case, which, reasoning deductively, supports the data at hand.

2 Mathematical probability

Given key concepts and relationships, inferential objectives and data, how should reasoning proceed? How can we characterize the nature and force of persuasion a mass of data conveys about a target inference? Workers in every field have had to address these questions as they arise with the kinds of inference and the kinds of evidence they normally address. Historically, the quest for principles of inference at a level that might transcend the particulars of fields and problems has received most attention in the fields of probability and statistics (unsurprisingly), philosophy and jurisprudence. Our interest is in the first of these and, in particular, mathematical or Pascalian (after Blaise Pascal) probability. For our purposes, the essential elements are a specified space of outcomes, or sample space; a parameter space; and a function that specifies the probabilities of outcomes given parameters, where probabilities are numbers between 0 and 1 that correspond to strength of belief and follow a few simple rules of combinations for 'events', where a 'Pascalian event' is a subset of the sample space. It is portentous that, given

parameter values, we can express the relative likeliness of a Pascalian event as compared to any other events; and given an event, we can express the relative likeliness of a given parameter value as compared to any other parameter value.

When it is possible to map the salient elements of an inferential problem into the framework of mathematical probability, powerful tools become available to combine explicitly the evidence that various probans (elements of evidence or intermediate conjectures) convey about probanda (target conjectures), as to both weight and direction of probative force. Inferential subtleties, such as chains of inferences, missingness, disparateness of sources of evidence and complexities of inter-relationships among probans and probanda, can be resolved. A properly structured statistical model embodies the salient qualitative patterns in the application at hand, and spells out, within that framework, the relationship between conjectures and evidence. It overlays a substantive model for the situation with a model for our knowledge of the situation, so that we may characterize and communicate what we come to believe – as to both content and conviction – and why we believe it – as to our assumptions, our conjectures, our evidence and the structure of our reasoning.

Perhaps the two most important building blocks are conditional independence and Bayes Theorem. Conditional independence is a tool for mapping Greeno's 'generative principles of a domain' into the framework of mathematical probability, expressing the substantive theory upon which deductive reasoning in a field is, and must be, based. This accomplished, Bayes Theorem is a tool for reversing the flow of reasoning – inductively, from observations to the more fundamental concepts of the domain, through these same structures, to expressions of revised belief in the language of mathematical probability.

3 *Conditional independence*

Two random variables x and y and *independent* if their joint probability distribution $p(x,y)$ is simply the product of their individual distributions – $p(x,y) = p(x)p(y)$. These variables are unrelated, in the sense that knowing the value of one provides no information about what the value of the other might be. Conditionally independent variables seem to be related – $p(x,y) \neq p(x)p(y)$ – but their co-occurrence can be understood as determined by the values of one or more other variables – $p(x,y|z) = p(x|z)p(y|z)$, where the conditional probability distribution $p(x|z)$ is the distribution of values of x, given the value z of another variable. The conjunction of sneezing, watery eyes and a runny nose described as a 'histemic

reaction' could be triggered by various causes such as an allergy or a cold; the specific symptoms play the role of *x*s and *y*s, while the status of reaction-causing conditions plays the role of *z*. The paradigms of a field supply 'explanations' of phenomena in terms of concepts, variables and putative conditional independence relationships. Judah Pearl (1988: 44) argues that inventing intervening variables is not merely a technical convenience but a natural element in human reasoning:

> [C]onditional independence is not a grace of nature for which we must wait passively, but rather a psychological necessity which we satisfy actively by organizing our knowledge in a specific way. An important tool in such organization is the identification of intermediate variables that induce conditional independence among observables; if such variables are not in our vocabulary, we create them. In medical diagnosis, for instance, when some symptoms directly influence one another, the medical profession invents a name for that interactions (e.g., 'syndrome,' 'complication,' 'pathological state') and treats it as a new auxiliary variable that induces conditional independence; dependency between any two interacting systems is fully attributed to the dependencies of each on the auxiliary variable.

In educational assessment, the variables in the student-competence model play the role of explanatory variables. They constitute the more abstract space in which we attempt to understand students' actions, evaluate their developing competences and plan further instruction. From the point of view of mathematical probability, the starting point for assessment is deductive reasoning through such a framework: 'How likely is a particular observation, from each of the possible values in the competence model?' The answer – the 'likelihood function' induced by this particular possible response – conveys the information that the observation conveys about competence, in the way competence is being conceived. If the observation is equally likely from students at all values of the variables in the competence model, it carries no information for inferences about those variables. If it is likely at some values but not others, it sways our belief in those directions, with strength in proportion to how much more likely the observation is at those values.

To illustrate this deductive stage of reasoning, we will use a student model based on the ACTFL reading guidelines. We will work with three collapsed levels of reading proficiency, namely, novice, intermediate and advanced, and map out the evidential grounding of two reading tasks, a multiple-choice question that is simply right or wrong and an extended performance task that supports four distinguishable levels of performance. We will assume for the moment that the requirements of background knowledge can be neglected. (This is *not* the case in many performance assessment

tasks, and we shall discuss how to extend the framework to deal with this in the following section on 'contextual dependencies.'

For each of the four reading competence categories, a panel in Figure 1 shows the probabilities of the different performance levels on the extended task. Each rectangle is a variable, with the probabilities associated with its different possibilities represented by bars that add up to one. Dashed bars represent certain knowledge – in Figure 1, looking at probabilistic expectations of responses if student competence level were known for a fact. The directed arrow in this so-called 'directed acyclic graph' (DAG) indicates the flow of deductive reasoning. We see that students at higher ACTFL levels are increasingly likely to do well on this task, although there is some chance for even advanced students to fare poorly and for novices to score well; that is, even knowing ACTFL with certainty would not give us perfect predictions of response. This is reasoning *from* an abstract conception of competence *to* expected performance – the 'forward reasoning' Holmes described to Watson. We determine these probabilities through theory, expert judgement, model-fitting (e.g., a latent class or item-response theory model), empirical data-gathering (e.g., observations on groups of students ascertained from external information to function at each of the three levels) or some combination (Andreassen *et al.*, 1987, illustrate these considerations in the context of medical diagnosis). Figure 2 shows similar conditional probabilities for the multiple-choice task. This hypothetical item is relatively easy, so we see in Figure 2 that only the novices will probably miss it. Intermediate students have 85% chances of getting it right and advanced students have 95% chances.

4 Bayes Theorem

We must reason inductively in most practical applications. In the language-task example, we will observe a student's performances in order to increase our knowledge about a student's level of competence on the ACTFL scale. When we can satisfactorily explicate the probabilities of observations given (inherently unobservable) values of variables in the student model as was illustrated above, Bayes Theorem provides a mechanism for reversing the flow of reasoning in a coherent manner. The mathematics of Bayes Theorem can be found in any statistical text; its central role in cognitive diagnosis and educational assessment is discussed more fully in Mislevy (1994; 1995). The essential idea is as follows:

- Before seeing observations, our belief about possible values of variables in the student model is expressed as a probability distribution – the *prior distribution.*

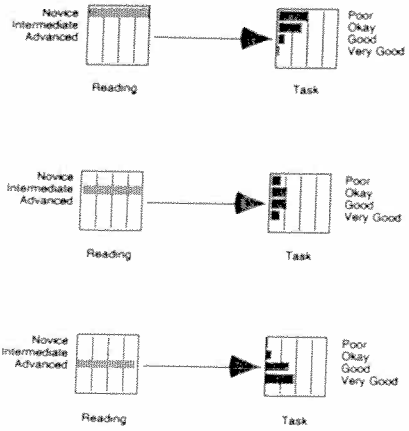


Figure 1 Conditional probabilities of extended-performance task responses, given competence level (deductive reasoning: three ACTFL levels, four levels of performance)
Notes: Nodes represent variables. Bars represent probabilities of potential values of a variable, adding up to one. A dashed bar represents certainty

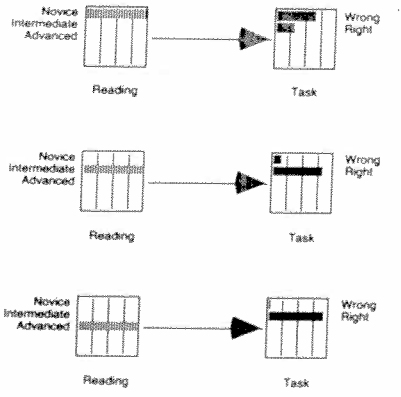


Figure 2 Conditional probabilities of multiple-choice task responses, given competence level (deductive reasoning: three ACTFL levels, right/wrong performance)
Notes: Nodes represent variables. Bars represent probabilities of potential values of a variable, adding up to one. A dashed bar represents certainty

- A particular value of an observable variable provides evidence about those values, in proportion to its probability of occurrence under each – the *likelihood function.*
- The product of the prior distribution and the likelihood function yield, for each possible value in the student model, a value proportional to its probability in a new distribution that reflects our revised beliefs – the *posterior distribution.*

Figure 3 represents inductive reasoning with the extended performance task. Inference flows in the opposite direction of the relationships represented by the directed arrow, which constitute the theory-driven structure of deductive reasoning – Holmes's 'backwards reasoning'. Now values of task performance become known with certainty when they are observed, and beliefs about possible values in the student model are updated. Each panels depicts the posterior probabilities for student competence induced by observing one of the four possible performance levels, starting from a prior distribution that considered the three levels equally likely. (In this special case, the posterior distribution is proportional to the likelihood function.) We see that, as would be expected, higher levels of observed performance shift our beliefs about students towards higher levels of competence. Figure 4 shows similar results for the multiple-choice task. Because this item is easy, a wrong response shifts our belief sharply towards a student being a novice, while a right response shifts belief away from novice, but does not provide much information to distinguish between intermediate and advanced.

5 *Bayesian inference networks*

Carrying out probability-based inference efficiently in complex networks of interdependent variables is an active topic in statistical research, spurred by applications in such diverse areas as forecasting, pedigree analysis, troubleshooting and medical diagnosis. Interest centres on obtaining the distributions of selected variables conditional on observed values of other variables, such as likely characteristics of offspring of selected animals given characteristics of their ancestors, or probabilities of disease states given symptoms and test results. The conditional independence relationships suggested by substantive theory play a central role in the topology of the network of inter-relationships in a system of variables. If the topology is favourable, such calculations can be carried out efficiently through generalizations of Bayes Theorem even in very large systems, by means of strictly local operations on small subsets of inter-related variables ('cliques') and their intersections. Discus-

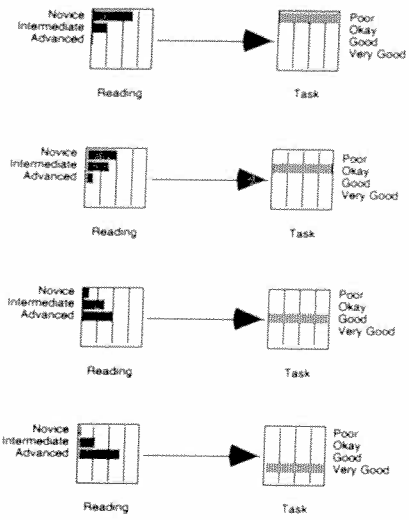


Figure 3 Posterior probabilities of competence levels, after observing extended-performance task response (inductive reasoning: three ACTFL levels, four levels of performance)
Notes: Nodes represent variables. Bars represent probabilities of potential values of a variable, adding up to one. A dashed bar represents certainty

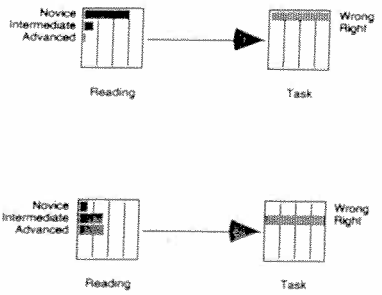


Figure 4 Posterior probabilities of competence levels, after observing multiple-choice task responses (inductive reasoning: three ACTFL levels, right/wrong performance)
Notes: Nodes represent variables. Bars represent probabilities of potential values of a variable, adding up to one. A dashed bar represents certainty

sions of construction and local computation in such Bayesian inference networks can be found in the statistical and expert-systems literature (see, for example, Lauritzen and Spiegelhalter, 1988; Shafer and Shenoy, 1988; computer programs that carry out the required computations include Andersen *et al.*, 1989, and Noetic Systems, 1991).

Figure 5 is a DAG for a simple inference network that combines the multiple-choice and extended-performance tasks introduced above. The three panels depict how belief about a student's level of competence is updated as the two responses are observed in turn. Directed arrows run from the student-model competence variable to each of the tasks, but there is no direct connection between the two; this indicates that they are conditionally independent given level of competence. It is in establishing such relationships that substantive theory comes into play: in defining unobservable variables that characterize students' state or structure of understanding, and observable variables that will convey evidence about that understanding; and in defining intervening variables and conditional independences through which deductive reasoning flows, so as to capture important substantive relationships and simplify computations. Note again the distinction between those assessment variables that are potentially observable and 'student-model variables' that are not, but in terms of which theories of knowledge and learning are framed (Mislevy, 1995).

The following sections extend our running ACTFL example in two ways in order to illustrate inferences about language competence that take into account the role of context and background in language acquisition and of observing more complex performances that require multiple aspects of competence. The focus is on the way this knowledge about the kind of competence we wish to make inferences about, and the way that it is manifest in complex settings, can be dealt with using probability-based inference.

V Dealing with context and situation

> [I]t appears that research on the measurement of the intellectual abilities generally associated with the term intelligence reached a point of diminishing returns a number of decades ago; though there has been continuing refinement of technical methods for test construction, progress has remained essentially asymptotic with regard to problems of predicting intellectual functioning outside of testing situations. An important reason suggested by the present analysis is continuing overdependence on the concept of context-free ability tests and consequent lack of analysis of the interactions and contexts (Estes, 1981: 18–19).

The 'traits' that achievement tests purportedly measure, such as

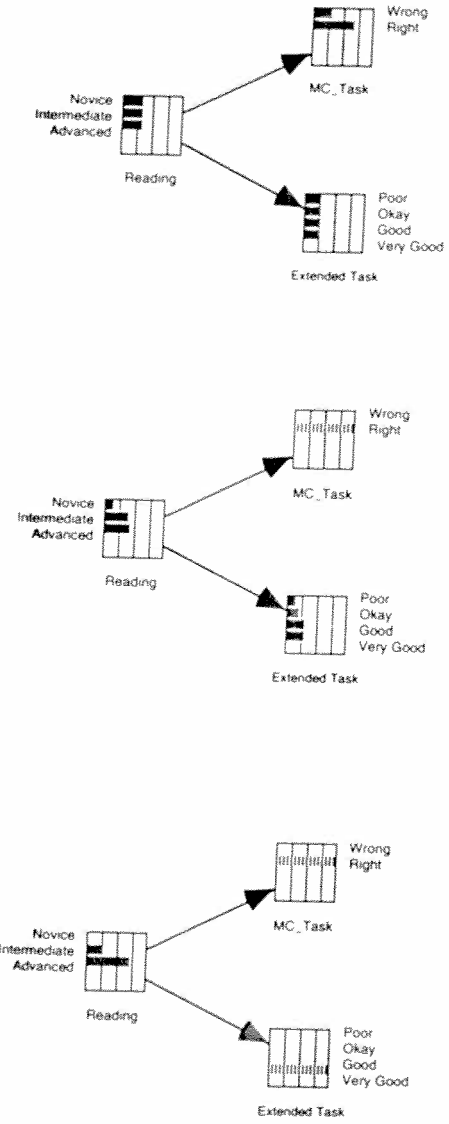


Figure 5 Successive updating of belief about competence level, after observing multiple-choice, then extended-performance, task results: (a) belief prior to observing any responses; (b) belief after observing a correct multiple-choice response; (c) belief after observing a correct multiple-choice response and a 'very good' extended-performance response

'mathematical ability', 'reading level' or 'physics achievement', do not exist *per se*. While test scores do tell us something about what students know and can do, any assessment task stimulates a unique constellation of knowledge, skill, strategies and motivation within each examinee. To some extent in any assessment comprising multiple tasks, which ones are relatively hard for some students are relatively easy for others, depending on the degree to which the tasks relate to the knowledge structures that students have, each in their own way, constructed. From the domain-behaviour perspective, this is 'noise', or measurement error. It obscures what one is interested in, namely, locating people along a single dimension as to a *general* behavioural tendency, and tasks that don't line up people in the same way are less informative than ones that do.

From the cognitive/developmental perspective, however, these interactions are fully expected, since knowledge typically develops first in context, then is extended and decontextualized so that it can be applied more broadly to other contexts. A given task may thus have the potential of providing considerable information about a given student, or none at all. Standard test theory does not address this concern at the level of tasks, but at the level of the combined test scores only after averaging results over multiple tasks; this is the issue of 'test validity' (Messick, 1989). But the greater investment each task requires and the more contextual knowledge it demands, the less efficient this approach becomes; hence the so-called 'low generalizability' problem some writers have attributed to performance assessments (e.g., Shavelson, Baxter and Pine, 1992). The in-depth project on proportionality that provides solid assessment information and a meaningful learning experience for the students whose prior knowledge structures it dovetails, becomes an unconscionable waste of time for students for whom it has no connection. The alternative is to take contextual and/or situational data into account when determining the evidential value that tasks provide about students' competencies. Practical assessment methods for doing this are discussed below. First, however, we illustrate the inferential situation with an extended inference network.

The mile posts outlined in the ACTFL reading guidelines are based on empirical evidence and theories about how competence in acquiring information from text in a foreign language develops. We have noted the contrast between intermediate readers' competence with texts '... about which the reader has personal interest or knowledge' with advanced readers' comprehension of '... texts which treat unfamiliar topics and situation' – a distinction fundamental to the underlying conception of developing language proficiency, which can alter the evidential import of the same behaviour

from the two students about their ACTFL levels. These relationships can be incorporated into a Bayesian inference network by extending the structure beyond nodes that characterize the situation only from an 'objective' point of view that pertains equally to all students – to nodes that vary across students in connection with their particular points of view; for example, whether a student has read a book upon which a reading passage is based. Consider an inference network that extends the one shown in Figure 1 by adding a new contextual variable, namely, whether the student is familiar or unfamiliar with the book in question.

Figure 6 illustrates expectations about performance as a function of given values of competence level and context familiarity, or the by-now familiar flow of deductive reasoning. Note the different expectations when the student is and is not familiar. Even students in the advanced category rarely perform well when they are unfamiliar with the context. When level of familiarity is not known, the expectations are an average of the two known conditions, and consequently much more diffuse. (The average is weighted by the proportion of students in each category who are and are not familiar with the book; for simplicity, this figure and the next assume a 50–50 split.) Figure 7 shows the results of inductive reasoning from observing a fairly low performance or a fairly high performance, under the conditions that we either 1) *know* the student is familiar; 2) *know* the student is *not* familiar; and 3) *don't know* whether the student is familiar. Note that the task conveys much more evidence about reading competence when we know the student is familiar with the context. That is, for a given level of observed performance, a more concentrated probability distribution, or a sharper inference, is obtained for level of proficiency if we know that the student is familiar with the context than if we know he or she is not, or if we don't know whether or not he or she is familiar. When low performance is observed in the third column where we don't know if the context is familiar to the student, appreciable probability remains that the student is intermediate or advanced; this is because both alternative explanations for low performance (low competence, and high competence but unfamiliar context) must be maintained.

Standard test theory for domain-behaviour inferences faces the third situation illustrated above. There are two standard test-theory methods for handling context dependency interaction between students and tasks in a domain: minimize it as much as possible, then average over whatever interaction remains with as many tasks as feasible. Minimizing it is accomplished by using tasks with which all examinees are similarly familiar or similarly unfamiliar. The costs

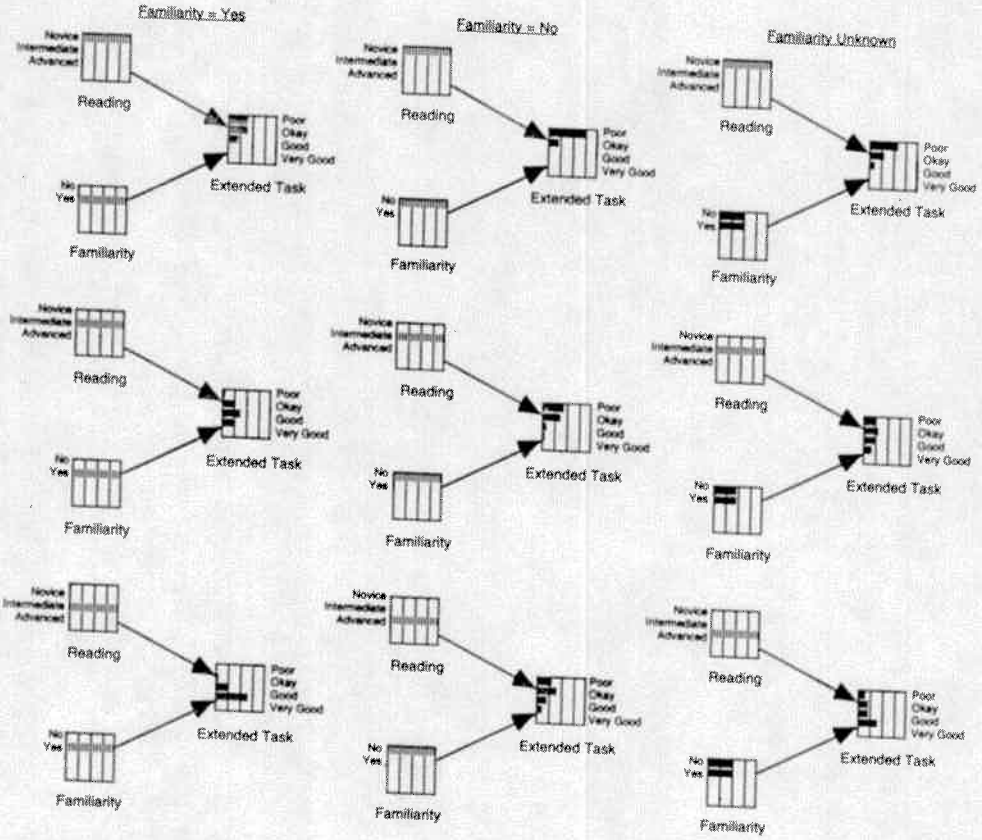


Figure 6 Conditional probabilities of extended-task performance, given competence level and task familiarity

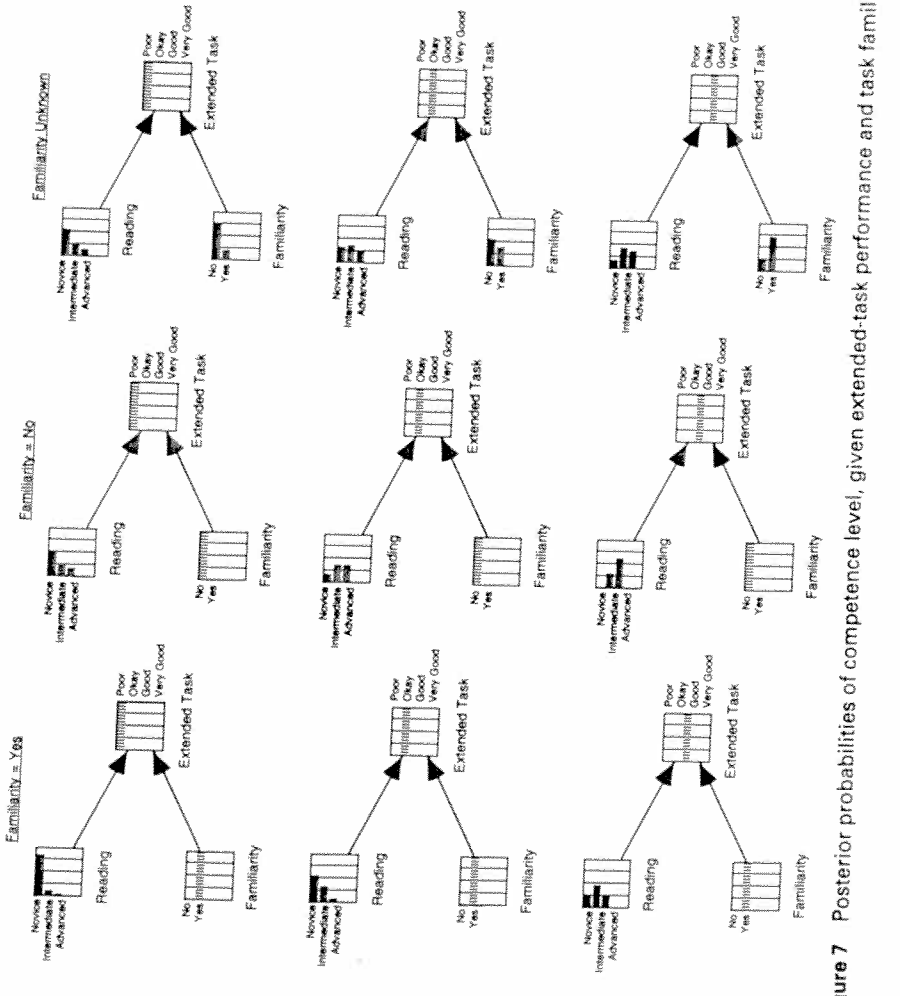


Figure 7 Posterior probabilities of competence level, given extended-task performance and task familiarity

are 1) avoiding tasks with which students may be personally interested, acquainted and able to display competences; and 2) making inferential errors of over- or underestimation of competence with respect to students for whom a particular task is atypically familiar or unfamiliar. Obviously the fewer tasks a student is administered, the more likely it is that this latter error occurs; therefore, averaging over as many tasks as possible helps to mitigate this problem. And it is an effective strategy with short, distinct, tasks such as multiple-choice items. It is less effective as each task becomes more time consuming.

Two alternative ways of handling contextual and situational effects both attempt to move from the last column in Figure 7 to the first or second column – preferably the first because that is where evidential value is highest, but at least if you know you're in the second column, you can use this information appropriately! The first way is to obtain contextual and situational data from each student along with task performance data. To the extent possible, findings about background variables are entered in an inference network along with task responses, and the conditional relationships among background and performance are taken into account. This strategy is taken in large-scale educational surveys such as the international assessments of mathematics in the form of 'opportunity to learn' measures (Platt, 1975). It is not effective for assessing individuals because tasks are administered without regard to these effects. This is analogous to administering a large battery of unrelated diagnostic tests to a hospital patient before we have any idea what the problem is, then only later trying to sort out which ones were meaningful ('turns out he has a broken leg, so I suppose we don't need any data from this CAT scan of his brain').

A second strategy is adapting what one observes to the student in accordance with values on what corresponds in our simple example to 'familiarity'. This can be done either by the assessor, as when an interviewer determines a subject of interest about which a conversation with a student can profitably take place, or by the student, as when choice among topics or exercises is provided. This is analogous in medical diagnosis to administering diagnostic tests sequentially, in the light of previous results and improved conjectures, and to asking the patient to provide information about what hurts and what happened. The choice strategy for educational assessment is most likely to provide interpretable evidence of competence if, no matter what the choice, evidence must be provided about the same more generally described competence, and it is made clear to the examinee what it desired and how it will be evaluated. Myford and Mislevy (1995) and Mislevy (1995) discuss how this strategy is

implemented and monitored in the College Entrance Examination Board's Advanced Placement Studio Art portfolio assessment.

VI Complex interaction of skills within tasks

Resnick and Resnick (1989) argue persuasively against the decontextualized and decomposed assessment tasks that characterize standard achievement tests. Genuine expertise, they claim, is contextualized and calls upon multiple aspects of skill and knowledge in concert. If this is what we seek to develop in students, should not they learn and be assessed in like terms to a far greater extent than they typically are? Creating assessment tasks that tap meaningful learning in engaging and effective ways is a significant challenge, but there are signs of progress (see, e.g., Lesh and Lamon, 1992). There has been less progress in figuring out just what to do with the 'data' that one obtains when students perform the tasks, both as to identifying just what is meaningful and how they are to be evaluated, and as to combining results across multiple and diverse tasks. This section addresses the latter problem in the framework of Bayesian inference networks; the former problem is discussed, among other places, in Myford and Mislevy (1995).

Consider again the ACTFL guidelines for reading, writing, speaking and listening. Suppose we want to assess students' competencies in Spanish in terms of these guidelines by means of the four tasks listed below. Figure 8 depicts the structure of the evidential relationships, showing baseline proportions of competence levels and task performances in a population of interest – our state of knowledge about a student from this population before we see any of his or her performances. The connections among the aspects of competence reflect the possibility of empirical relationships among them in a population of interest (e.g., people who can write well in a foreign language might usually read well; a weaker relationship may exist between writing and listening):

- *Task A* is the extended-performance reading task introduced above, providing a bit of direct evidence about reading only.[2]

[2] Direct evidence about reading competence may provide *in*direct evidence about other competencies, to the extent that people who tend to do well in one aspect of language competence tend to do well in others. But the four-aspect ACTFL guidelines already embody the results' research on this topic: there are more finely detailed aspects of competence within reading that *do* tend to develop together, and are thus subsumed in the more generally defined reading guidelines; the same holds for listening, writing and speaking. This finer breakdown would in fact be required in instruction. Competencies in the four main aspects, however, are seen to follow very different paths in different people. Graduate students may be required to learn to read a foreign language, for example, but acquire few listening or speaking skills. Conversely, extended visitors to a foreign country may pick up speaking and listening skills rapidly with only reading or writing skills.

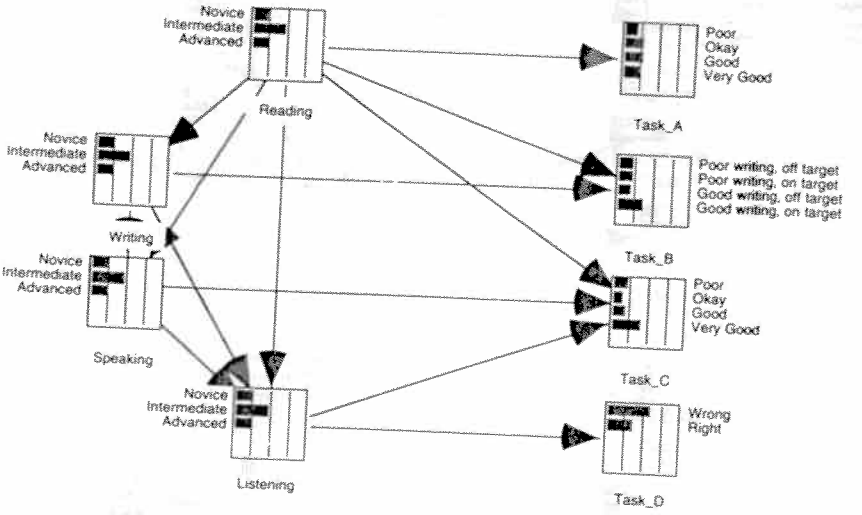


Figure 8 Evidential structure of four tasks and four aspects of competence (status of belief before observing any responses)

The relationship between Task A and reading competence is the one shown in Figure 1, but now embedded in a larger context.

- *Task B* is reading a complex passage and writing a response to a question about it. It is possible to obtain evidence about both reading and writing, but a dependency must be accounted for: low levels of writing competence eliminate the chance to acquire direct evidence about reading. A sensible response competently written provides evidence about higher competence about both reading and writing (the first panel of Figure 9). A well written but off-task response shifts belief towards higher competence in writing but lower levels of competence in reading (the second panel of Figure 9). A poorly written and off-target response shifts belief away from higher levels of both reading and writing (the final panel of Figure 9).
- *Task C* asks the student to listen to a taped conversation with a transcript provided, then talk about the interaction. A well spoken and accurate response signifies higher speaking competence (see the first panel of Figure 10), and shifts beliefs about both listening and speaking higher – though not for either as much as for speaking, since we don't know whether the student listened to the conversation, read the transcript or both. An 'okay' response shifts beliefs about speaking towards 'intermediate', and both listening and reading in the same direction – though again not as strongly because of the multiple explanations for this observation (the second panel of Figure 10). A

'poor' response shifts belief about all three aspects of competence involved in the tasks downward. Possible causes, the situations of which are averaged over in the result, include failure at the stage of understanding the message – i.e., lack of both listening and reading skills – and/or the stage of responding – i.e., low speaking skills (the final panel of Figure 10).

- *Task D* asks the student to listen to a taped conversation, and indicate by raising his or her hand when a business transaction is completed. Direct evidence about only listening competence is obtained. Figure 11 shows the results of observing a student respond correctly to Task D and do well on Task A after having done poorly on Task C. That is, the final panel of Figure 10 was the state of belief before observing this new correct response to Task D. Obtaining evidence that the student may have both reading and listening helps sort out the possibilities that could have led to poor performance in Task C; it is now more likely that speaking competence was the source of difficulty there.

For the reasons discussed above, I do not generally favour having holistic quality standards applied uniquely to individual tasks, each of which probes different mixtures of aspects of competence. The combination of idiosyncratic scores by any such means cannot capture differences among configurations of competence, and ignores patterns of strength or weakness among aspects of competence across tasks. The meaning of combined idiosyncratic scores is unambiguous only when almost all performances are successful or almost all are unsuccessful. I much prefer a structure under which evidence about various aspects of competence evinced by a task are evaluated in the light of their mixture, accounting for their interdependencies. Having coherently interpreted evidence about aspects of competencies, one can then collapse this information in various ways for summarization, reporting and evaluation. (See Haertel, 1989, and Haertel and Wiley, 1993, on the topic of explicating evidential structure of performance tasks.)

VII Conclusion

We do not build probability models for most of the reasoning we do, either in our jobs or our everyday lives. We continually reason deductively, inductively and abductively, to be sure, but not through explicit formal models. Why not? Partly because we use heuristics, which, though suboptimal (e.g., Kahneman, Slovic and Tversky, 1982), generally suffice for our purposes; more importantly, because much of our reasoning concerns domains we know something about. Attending to the right features of a situation and reasoning through

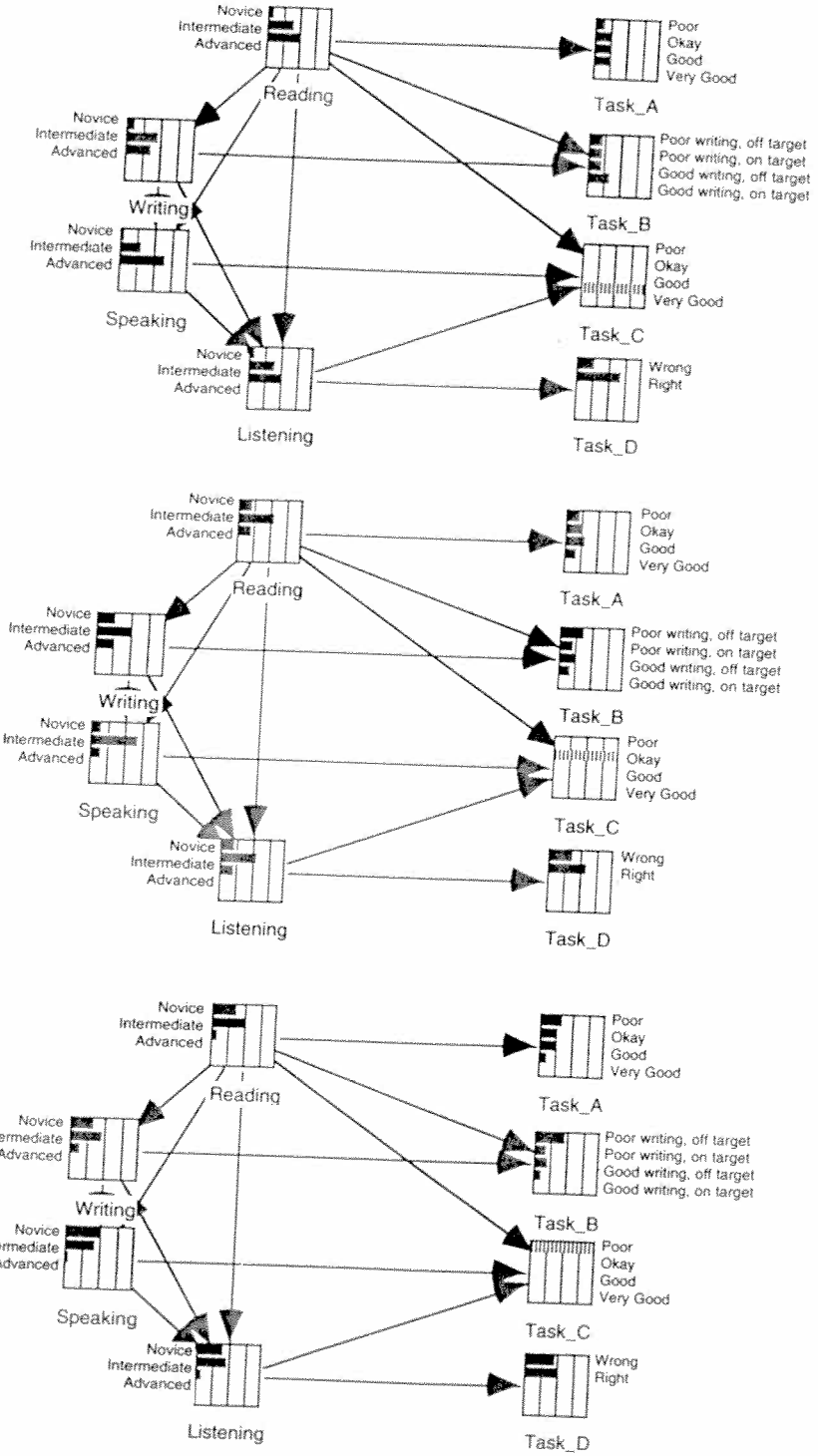


Figure 10 Posterior probabilities for competences, after observing various Task C responses

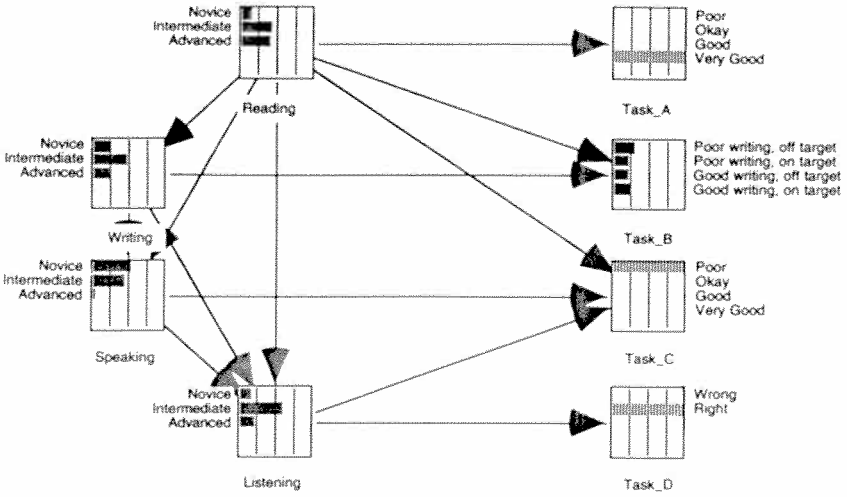


Figure 11 Posterior probabilities for competences, after observing a poor task C response, a very good Task A response and a correct Task D response

the right relationships, informally or even unconsciously, provides some robustness against suboptimal use of available information within that structure. Heuristics, habits, rules of thumb, standards of proof and typical operating procedures guide practice in substantive domains, more or less in response to what seems to have worked in the past and what seems to have led to trouble. This inferential machinery coevolves with, and is intimately intertwined with, the problems, the concepts, the constraints and the methodologies of the field (Kuhn, 1970: 109). But difficulties arise when inferential problems become so complex that the usual heuristics fail, when the costs of unexamined standard practices become exorbitant or when novel problems appear. It is in these situations that more generally framed and formally developed systems of inference provide their greatest value.

We face this situation today in language-learning assessment; indeed, in educational assessment in general. The standard methods, rules of thumb and canons of good practice have evolved to address inference in a universe of discourse more restricted with respect to generative principles and observational material than the one that now commands our attention. To support inference in this extended universe of discourse about assessment, we will simply have to work through many problems from first principles. We must figure out just what it is we want to make inferences about – that is, first aspects, then models, of student competence. We must learn to construct situations that evoke evidence about these. We must

explicate the probabilistic structure between the nonobservable constructs and observations. We must (as is the focus of this article) use analytical methods that characterize the import and weight of evidence for our inferences. Sometimes this will be standard, familiar test theory, such as classical test theory, item response or factor analysis. Sometimes it will not be. But probability-based inference can be gainfully applied to attack many of these problems, if not always with off-the-shelf tools. The first order of business for those of us in test theory, therefore, is to develop conceptual framework and analytic tools for carrying out these studies.

VIII References

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

Andersen, S.K., Jensen, F.V., Olesen, K.G. and **Jensen, F.** 1989: *HUGIN: a shell for building Bayesian belief universes for expert systems* (computer program). Aalborg, Denmark: HUGIN Expert.

Andreassen, S., Woldbye, M., Falck, B. and **Andersen, S.K.** 1987: MUNIN: a causal probabilistic network for interpretation of electromyographic findings. In *Proceedings of the 10th International Joint Conference on Artificial Intelligence.* Milan: Kaufmann, 366–72.

Doyle, A.C. 1930: *The complete works of Sherlock Holmes.* New York: Doubleday.

Estes, W.K. 1981: Intelligence and learning. In Friedman, M.P., Das, J.P. and O'Connor, N., editors, *Intelligence and learning.* New York: Plenum, 3–23.

Glaser, R. 1991: Expertise and assessment. In Wittrock, M.C. and Baker, E.L., editors, *Testing and cognition.* Englewood Cliffs, NJ: Prentice-Hall, 17–30.

Greeno, J.G. 1989: A perspective on thinking. *American Psychologist* 44, 134–41.

Haertel, E.H. 1989: Using restricted latent class models to map the skill structure of achievement test items. *Journal of Educational Measurement* 26, 301–21.

Haertel, E.H. and **Wiley, D.E.** 1993: Representations of ability structures: implications for testing. In Frederiksen, N., Mislevy, R.J. and Bejar, I.I., editors, *Test theory for a new generation of tests.* Hillsdale, NJ: Lawrence Erlbaum Associates, 359–84.

Kahneman, D., Slovic, P. and **Tversky, A.** 1982: *Judgment under uncertainty: heuristics and biases.* Cambridge: Cambridge University Press.

Krathwohl, D.R. and **Payne, D.A.** 1971: Defining and assessing educational objectives. In Thorndike, R.L., editor, *Educational measurement* (2nd edn). Washington, DC: American Council on Education, 17–45.

Kuhn, T.S. 1970: *The structure of scientific revolutions* (2nd edn). Chicago, IL: University of Chicago Press.

Lauritzen, S.L. and **Spiegelhalter, D.J.** 1988: Local computations with

probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society, Series B* 50, 157–224.

Lesh, R.A. and **Lamon, S.**, editors, 1992: *Assessments of authentic performance in school mathematics.* Washington, DC: American Association for the Advancement of Science.

Messick, S. 1989: Validity. In Linn, R.L., editor, *Educational measurement* (3rd edn). New York: American Council on Education/Macmillan, 13–103.

Mislevy, J.R.J. 1994: Evidence and inference in educational assessment (1994 presidential address to the Psychometric Society). *Psychometrika* 59, 439–83.

Mislevy, R.J. 1995: Probability-based inference in cognitive diagnosis. In Nichols, P., Chipman, S. and Brennan, R., editors, *Cognitively diagnostic assessment.* Hillsdale, NJ: Lawrence Erlbaum Associates, 43–71.

Myford, C.M. and **Mislevy, R.J.** 1995: *Monitoring and improving a portfolio assessment system.* Center for Performance Assessment Research Report. Princeton, NJ: Center for Performance Assessment, Educational Testing Service.

Noetic Systems Inc. 1991: *ERGO* (computer program). Baltimore, MD: Noetic Systems Inc.

Pearl, J. 1988: *Probabilistic reasoning in intelligent systems: networks of plausible inference.* San Mateo, CA: Kaufmann.

Platt, W.J. 1975: Policy making and international studies in educational evaluation. In Purves, A.C. and Levine, D.U., editors, *Educational policy and international assessment.* Berkeley, CA: McCutchen, 33–59.

Resnick, L.B. and **Resnick, D.P.** 1989: Assessing the thinking curriculum: new tools for educational reform. In Gifford, B.R. and O'Conner, M.C., editors, *Future assessments: changing views of aptitude, achievement, and instruction.* Boston, MA: Kluwer, 37–75.

Schum, D.A. 1987: *Evidence and inference for the intelligence analyst.* Lanham, MD: University Press of America.

Shafer, G. and **Shenoy, P.** 1988: *Bayesian and belief-function propagation. Working Paper* 121. Lawrence, KS: School of Business, University of Kansas.

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

Stake, R.E. 1991: The teacher, standardized testing, and prospects of revolution. *Phi Delta Kappan* 73, 243–47.