4

Developing and Validating Cognitive Models in Assessment

Madeleine Keehner, Joanna S. Gorin, Gary Feng, and Irvin R. Katz

Definition of Cognitive Models

A **cognitive model** is a theoretical account of the processes and steps assumed to occur during complex cognitive phenomena, such as problem solving, decision making, planning, or memory retrieval (Busemeyer & Diederich, 2010; Markman, 1999). Cognitive models are a formal attempt to characterize these processes, and a model makes particular assumptions about the mechanisms involved in the phenomenon it is describing. In the field of cognitive psychology, the general goal in developing a cognitive model is to account for observations from empirical studies and make detailed predictions about behavioral outcomes that would be expected when completing a specific task or activity under a particular set of conditions (Busemeyer & Diederich, 2010). For example, current models of semantic memory assume that category knowledge is stored in a *spreading-activation network*, in which the strength of a memory trace increases or decays according to how much it is activated. This kind of model superseded earlier *hierarchical models*, due to the weight of evidence from studies testing recall accuracy and response times under different experimental conditions (Anderson, 1983). Cognitive psychologists have developed and tested many such theoretical models to describe, explain, and predict basic cognitive phenomena (Busemeyer & Diederich, 2010).

Cognitive models also have potential utility in educational assessment. The underlying logic is that empirically supported cognitive models of **target constructs** can provide valid and useful a priori assumptions and principles for **item design**, scoring, and validation. As Leighton and Gierl (2007) state, the cognitive models we use in this domain often have to be broader in scope (incorporating students' knowledge and skills at different levels of learning) and more focused on education tasks than those typically developed in cognitive psychology. As these authors also point out, due to the

The Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications, First Edition.
Edited by André A. Rupp and Jacqueline P. Leighton.

realities and practicalities of large-scale assessment development, cognitive models for assessment may not always meet the same standards for evidentiary support and empirical testing of assumptions. Still, despite the greater challenges of gathering and analyzing evidence of cognition in this field, there are some sources of information about the **cognitive processes** of test-takers during assessment tasks that can be effectively pursued even in operational contexts. These are the focus of this chapter.

Having assessments grounded in cognitive models should increase the likelihood that scores from our tests reflect the constructs we are targeting and thus provide a strong evidentiary basis for a range of score-based decisions, interpretations, and uses (Bachman, 2005; Leighton & Gierl, 2007). To achieve this promise, two conditions must be satisfied: (1) we must have sufficient knowledge of the construct to have an appropriate cognitive model defined, and (2) we must have a means by which to evaluate the extent to which our test **items**, and the resulting scores, decisions, and inferences, are accurate reflections of that cognitive model. Other chapters in this volume address the first issue in terms of the different types of cognitive models and their utility in educational assessment. Our focus in this chapter is how one would go about developing and validating a cognitive model that can effectively support assessment goals. We hope that the tools and methods described provide some useful empirical avenues for generating and testing assumptions about cognition.

First, a note about terminology. In the sense that we are choosing to use the terms in this chapter, a cognitive model is not necessarily the same as a target construct. A target construct, such as expository writing skills, is what we hope to measure; that is, we devise a task to gather evidence from which we can make a (usually quantifiable) claim about a student's competency in that construct. In order to identify a construct as a measurement target, we need to define or describe the construct such that we can devise a task to elicit evidence that can be used to infer a student's competency in that construct. In principle, this could be done without a cognitive model. For example, an element of the taught curriculum could be identified as the target we hope to measure, without any formal model of the cognitive process involved in engaging in the relevant activities (see Leighton & Gierl, 2007).

By contrast, a cognitive model attempts to describe and/or explain the processes that are involved when we engage in a given cognitive activity. That activity could be a component of the taught curriculum, or a particular task designed to measure a given skill, or any other educational or everyday cognitive activity. Cognitive models may of course also be developed for purposes other than assessment, such as to support pedagogical decisions, or simply for the goal of developing and testing theory in and of itself.

In cognitive psychology, the ultimate goal is to be able to understand, describe, and predict behavior. Thus, the focus of scientific study and theory development is the cognitive process that underlies the behavioral responses, and the experimental tasks that are devised to elicit those responses are a means to that end, rather than an end in themselves (although, of course, the design of those tasks do shape the evidence collected and thus shape inferences about the constructs). In educational assessment, by contrast, the tasks we use to measure cognitive phenomena are of the utmost importance, since it is from these tasks that our scores are generated and ultimately used. A cognitive model provides a cognitively-grounded account of the processes involved in the construct we are targeting in our assessment. Therefore, in the educational

assessment context, the cognitive model is tied essentially to the task. The degree to which that task adequately represents the general behavior of interest is an important, but not identical, question as to whether we have properly specified the underlying cognition of our tasks. The initial question is about the quality of evidence from the task for making claims about test-takers and this can be informed by the cognitive model; a subsequent question is about the quality of the cognitive model itself.

In some previous discussions of cognitive models for educational assessment, a distinction has been made between two types of cognitive models, both of which are relevant to assessment validation (Ferrara et al., 2004; Gorin, 2006a). The first is the cognitive model of the target construct; that is, the generally defined set of cognitive processes, skills, and abilities that make up the construct we are aiming to measure (e.g., reading comprehension as a general process). This broadly characterized set of cognitive processes has sometimes been called the **intended construct** (Gorin, 2006a). Developing and validating a complete cognitive model of an intended construct is equivalent to theory development, and is closer to the fundamental description and explanation of cognitive processes sought by cognitive psychologists as part of the goals of that discipline. Although in principle these more general cognitive models include all of the cognitive processes that are relevant to the construct of interest, they are not typically the starting point for developing assessments.

Instead, when developing assessments, we need to create specific instantiations of the more general construct by means of test items and tasks (e.g., the reading comprehension and response processes engaged by a particular expository text and by a particular set of questions about that text). Based on the specific format, content, context, and administration conditions of a particular test, one could describe the cognitive processes that are associated with performance on these items by means of a more specific cognitive model; Gorin (2006a) refers to this more specific cognitive model as the **enacted construct**. Our primary focus in this chapter is the development and validation of cognitive models of enacted constructs (i.e., knowledge, skills, and abilities that are elicited by particular assessment tasks) since this is more typical of what assessment professionals might seek to do, and our goal here is to provide information that is relevant to and useful for the purposes of assessment. Thus, when we refer to cognitive models hereafter, we are generally assuming that the goal is to model an enacted construct, or, in other words, to describe and validate the cognitive processes involved in a particular assessment task.

Model Development and Validation Methods

Perhaps the most natural question for practitioners, most of whom are not cognitive psychologists, is how to begin to develop a cognitive model of the constructs enacted in their tasks. Theories of cognition, learning, expertise, training, and assessment in various domains can provide rich sources of information for model development. Figure 4.1 presents an iterative process of cognitive model development and validation, beginning with the broader model of the intended construct based on existing theory from relevant disciplines and finishing with an empirically validated model of the enacted construct relevant to the particular tasks and test scores.

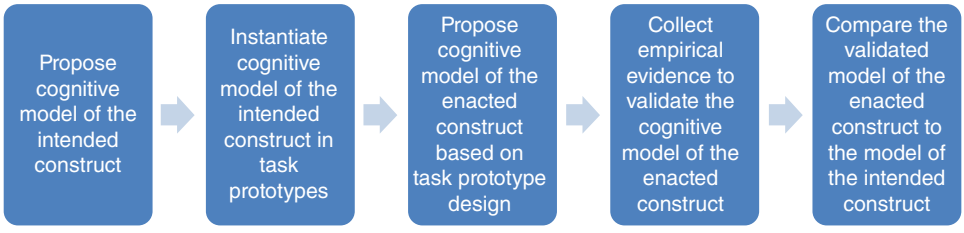

Figure 4.1 A development and validation process for cognitive models in assessment.

The general process outlined in Figure 4.1 is similar to the scientific process for building cognitive theories – propose a theory, generate observational contexts (i.e., tasks) to empirically test the theory, examine the observed results relative to the model's predicted results, and draw conclusions about the model based on the results. The difference, unique to assessment research, is that we focus on the task as the unit of analysis with the ultimate goal of a fully specified and validated **task model** (i.e., a model of the enacted construct) that aligns with the general cognitive model (i.e., the intended construct).

The amount, type, and quality of extant literature on the general intended construct will affect the number of iterations that are likely to be needed, and in some cases further development of an appropriate intended construct may be necessary. Furthermore, to the extent that the form of the assessment task is novel (i.e., not a form that has undergone extensive empirical research), multiple iterations will be needed to refine and understand task elements and how they cue associated processes in test takers as part of the enacted construct, relative to the intended construct.

A critical step in this process of model development is the identification, accumulation, and interpretation of evidence (i.e., data) to inform conclusions about the cognitive processes elicited by our tasks. Depending on the nature of the construct and the task, evidence can come from a range of sources. External or "offline" sources of **validity** are collected "outside" of the task, either temporally or physically. This category of evidence includes the kinds of data that are typically used to validate cognitive models using an exploration of **nomothetic span** (Embretson, 1983). External or offline evidence is typically easier to capture since a separate activity can be designed or adopted from established methods and separately administered and the data produced can require relatively straightforward analysis (e.g., analyses using correlation coefficients or mean differences).

By contrast, "online" evidence of validity is data collected while the test-taker is completing the task. Some online evidence might be argued to fall into the category of internal validity data in the sense described by Embretson (1983) such as keystroke logs that are presumed to reflect the target cognitive processes of text generation that occur during a writing task (e.g., Miller, 2000). However, online evidence can also include data that are external to the task itself but are still gathered concurrently with task completion such as EEG recordings or psychophysiological measures of arousal in test takers captured during an educational assessment. Methods that produce online evidence must be administered during – and be temporally aligned with – the steps in the assessment task in order to offer a window into, or be correlated with, cognitive events

as they happened during task completion. This is typically a more challenging goal since the methods must be adapted or designed from scratch to fit the task and must be able to be administered without disrupting or altering task performance. Because of these requirements, they may involve costly equipment or data collection methods and require relatively complex analyses in order to make sense of the complex, indirect, or sequential kinds of evidence they produce. Ultimately, most researchers would agree that a combination of online and offline evidence is likely to provide the most complete understanding of the cognition associated with a particular task.

Offline Evidence: Methods and Approaches

When validating a cognitive model for assessment goals, relevant measures include ratings or measures external to an individual item that could be compared to some summary indicator that represents the item's measurement/psychometric properties. Here we describe three offline approaches that appear in the assessment literature to validate cognitive models: (a) correlations among items' statistical properties and item attributes, (b) psychometric models relating item attributes and item response processes, and (c) experimental manipulations of items.

Correlations among item attributes, including item difficulty. Correlational evidence is frequently used as part of **construct validity** arguments (Cronbach & Meehl, 1955; Kane, 2008; Messick, 1989). Correlations among scores from measures, including the to-be-validated measure and other external measures, are often examined in **multitrait multimethod** matrices (Campbell & Fiske, 1959). Statistically significant correlations among theoretically related constructs provide convergent evidence while zero or near-zero correlations among theoretically unrelated constructs are also consistent with expected evidentiary patterns. Similarly, correlations among scores from the internal elements of a test (e.g., inter-item correlations) or internal elements within the entire measure (e.g., item-total correlations) provide evidence about the **internal structure** of a measure. When this evidence is evaluated relative to a theory about the dimensionality of the measure, it can be viewed as construct validity evidence.

When considering how to validate a cognitive model, one can take the analysis to a more fine-grained level by investigating the statistical relationships among item parameters and variables that reflect different processing components, skills, or aspects of required knowledge. One approach is through correlational analysis of item difficulty statistics, also called **item difficulty modeling**. An **item difficulty model** (IDM) is a cognitive model of an item that specifies the processes, skills, and/or knowledge required to solve that item, and the impact of each on the overall item difficulty (Gorin, 2006b). The key to item difficulty modeling is to identify the relevant features that drive item processing and to estimate their impact. Done well, it should give rise to a strong understanding of the intended construct as well as an understanding of how the construct has been instantiated (enacted) by a particular task.

The specification of an IDM thus begins with a specification of the hypothesized set of skills, knowledge, and processes required to respond correctly to an item. Starting from an examination of item features, a preliminary list is often generated of cognitive

processes that the item is expected to elicit. This list would typically be based on theoretical literature relevant to the content area and, if available, empirical investigations of information processing. For each cognitive process specified, one or more variables is created representing various item features that are presumed to be associated with that process.

For example, an item feature of a mathematics question might be the number of sequential components presented in the problem. A cognitive consequence of manipulating this feature might be the number of operations that must be applied to correctly solve the item. A formal representation of the model is often generated as an item-by-skill matrix called a **Q-matrix** (Tatsuoka, 1995). A Q-matrix contains a mathematical representation of the skills (expected cognitive processes) required by each item according to its design features. The cognitive processes might reflect lower-level atomistic steps in a solution or higher-level aggregate processes (e.g., "understand" or "check solution"; Katz, Martinez, Sheehan, & Tatsuoka, 1998). Continuous or discrete values can be used depending on the nature of the items and the skills they are expected to invoke in test-takers. In many cases, a dichotomous code (0/1) is used simply to indicate whether a particular skill is required for a correct item solution. Once a cognitive model of the task has been specified in this way, it can be validated.

The next step is to estimate the existence and strength of the relationship between the skills and *item operating characteristics*, including item difficulty. One of the most common approaches is to use correlational analyses. For example, item difficulty parameter estimates, obtained from operational administrations of test items, may be regressed on the quantified item features. Analogous to the convergent evidence we seek in a multitrait multimethod analysis, evidence supporting the cognitive model comes from a statistically significant explanation of item difficulty for each of the proposed item features, which was found to influence cognitive processing as expected, as well as the overall model fit for the set of features. The difficulty modeling process is iterative such that item features are added to or removed from the difficulty model based on their contribution to the explanatory power of the model. The ultimate goal is to develop a model that most completely accounts for item difficulty based on features of the test question that are associated with cognitive processes.

IDMs have been developed for a range of educational and psychological constructs, including abstract reasoning (Embretson, 1998), quantitative reasoning (Embretson, 2010; Enright, Morley, & Sheehan, 2002), verbal reasoning (Gorin & Embretson, 2006; Sheehan & Ginther, 2001), and spatial reasoning (Embretson & Gorin, 2001). In a conceptually related approach, research has been conducted to explore when and why items differ in difficulty when administered to linguistically and culturally diverse groups (Erickan et al., 2010). In this study, which used explanatory **differential item functioning** (DIF) analysis and data from **think-aloud protocols**, IDMs were essentially developed for conditional item parameter differences in order to validate a cognitive model, in that the source of the DIF was considered to be a function of differential cognitive processing associated with items for different sets of students. In general, these models have shown moderate to strong explanatory power accounting for anywhere from 45% to 70% of the variance in item difficulty parameter estimates, and have been used in applications such as **automatic item generation** and diagnostic score reporting.

Cognitive-psychometric modeling. There are also formal psychometric models that incorporate cognitive features directly into the mathematical formulation of the test score. The majority of these models, which are often termed **cognitive psychometric models** or singly explanatory models (Wilson, De Boeck, & Carstensen, 2008) have been introduced in response to criticisms that traditional **statistical models** are disconnected from cognitive theory. Although detailed description of these models is not the focus here, it is important to mention some examples at least briefly to provide some general context. These methods are *latent trait models* – formal attempts to statistically model non-observable traits (e.g., ability) using the variance in observable behaviors (e.g., student responses) on indicators designed to measure the target construct (e.g., assessment items). Such models are useful for substantive examinations of score meaning and cognitive model validity because they provide a mechanism for testing the fit of cognitive processing models to the data.

Three related models developed for use in cognitive psychometric modeling are the *linear logistic latent trait model* (LLTM) (Fischer, 1973), the *multicomponent latent trait model* (MLTM) (Whitely, 1980), and the *general multicomponent latent trait model* (GLTM) (Embretson, 1984). The unidimensional LLTM incorporates cognitive attributes into the calculation of probabilities of a correct response to an item. The model includes this information in the form of regression weights representing the impact that any one cognitive component of a trait may have on the difficulty of an item. Essentially, item difficulty is decomposed into a linear combination of cognitive attributes and the impact of those attributes on solving an item. The MLTM, a multidimensional extension of the LLTM, can be applied to items measuring traits with multiple components (Whitely, 1980). In this model, it is assumed that correct sequential completion of components must occur in order to respond correctly to the item. Failure to complete any of the components results in an incorrect response. Finally, the GLTM combines both the LLTM and the MLTM and can be used for complex data that neither of the other two methods can model effectively. The GLTM includes both components and complexity factors such as combining the explanatory breakdown of item parameters internally in the model with a multidimensional representation (Embretson, 1984).

Another set of models based on classification rules have also leveraged cognitive information in modeling assessment data. Applications of these models have primarily focused on diagnostic score reporting rather than cognitive model validation (although cognitive model validity might be viewed as a prerequisite for valid diagnostic score reporting). Tatsuoka's **rule-space method** (RSM) is an approach to data analysis designed to provide feedback to groups and individuals regarding skill mastery (Tatsuoka, 1985, 1995). It begins with an evaluation of skills needed to solve a problem correctly. The student's skill level is diagnosed based on responses to items and the association between the items and skills. The RSM has been successfully applied to tests of mathematics, reading comprehension, and listening, to generate cognitive score reports of student ability (Buck & Tatsuoka, 1998; Buck, Tatsuoka, & Kostin, 1997; Tatsuoka, Corta, & Guerrero, 2004).

Finally, there is a class of parametric model families that includes cognitive-psychometric models for diagnosis that are structured as constrained latent class models, where each latent class is associated with a different diagnostic state. These are exemplified by the

general diagnostic model (von Davier, 2011) as well as the *deterministic-input noisy-"and" gate model* (Junker & Sijtsma, 2001, cited in von Davier, 2011) and the *loglinear cognitive diagnosis model* (Henson, Templin, & Willse, 2009, cited in von Davier, 2011), which are all statistically very similar. To date, the general diagnostic model has been fitted to data from language competency assessments as well as large-scale NAEP data (von Davier, 2005; Xu & von Davier, 2006). The *fusion model* (Hartz, 2002) is also statistically very similar, and is one of the most parametrically flexible (but also one of the most difficult to estimate) models.

Experimental manipulations. In an educational assessment context, experimental methods can be a useful tool for validating cognitive models. Frederiksen (1986) relates the experimental method to construct validity quoting Messick's (1975, p. 995) statement that "test validation in the construct framework is integrated with hypothesis testing and with all the philosophical and empirical means by which scientific theories are evaluated." Bearing in mind the distinction between validating the cognitive model of the intended construct and that of the enacted construct, we would argue that, in this domain, experimental methods can be especially relevant for examining an enacted cognitive model.

Two general experimental designs are appropriate for this purpose: (1) manipulations of item features (Embretson & Gorin, 2001; Enright, Morely, & Sheehan, 2002; Gorin, 2005; Katz, Lipps, & Trafton, 2002), and (2) manipulations of item format/context (Katz, Bennett, & Berger, 2000; Katz & Lautenschlager, 1994, 2001; Powers & Wilson, 1993). In the first approach, experimenters manipulate features of items associated with expected cognitive processes such as the number of variables in a math problem and examine the effects of these manipulations on statistical item parameters. Manipulations of items that cause changes in the item parameter estimates are assumed to play a role in how test-takers cognitively process the item. Item manipulations that do not affect the item parameter estimates are assumed to be incidental (i.e., not important) to cognitive processing (Bejar et al., 2002).

The other type of experimental design, manipulation of item format or context, deals with how modifying the conditions under which a person responds to an item changes their cognitive processing and, thus, the item parameter estimates. In this second approach, experimenters manipulate factors that are associated with the administration of an item but do not make changes to the item itself. For example, the same reading comprehension item may be presented in two different conditions such as with and without an accompanying passage, with two versions of an accompanying passage, or at different time points within an assessment; similar conclusions can be made as with direct manipulation of item features. Changes to item format or context that affect the difficulty level of an item are presumed to affect cognitive processing.

Several experimental studies of reading comprehension items have been conducted in the service of validating elements of a cognitive model of the enacted construct. Similar to the IDM approach, researchers parsed sources of processing difficulty for a particular reading assessment task – passage-based **multiple choice** reading comprehension items – according to the components of items and evaluated their differential effects in the passage versus the questions (Katz & Lautenschlager, 1994, 2001; Powers & Wilson, 1993). One experimental study was devised in which

participants responded to the same set of questions either with or without the associated passages. Results showed that the difficulty of the items did not differ significantly when administered with or without the passage (Katz & Lautenschlager, 1994). These studies with college-aged students demonstrated that items from secondary and post-secondary achievement tests could, in fact, be solved without reading the passage associated with the question. The results challenged the previously assumed alignment of the cognitive model of the enacted construct with that of the intended construct. That is, theoretical models of reading comprehension include cognitive processes associated with the encoding of text and the construction of a mental representation of the text that is used to respond to the test questions (Kintsch & vanDijk, 1978), yet the experimental results suggested that the enacted construct in this case did not include some of these critical cognitive processes.

Gorin (2005) also demonstrated the value of experimental methods for validating cognitive model components. In her study of GRE reading comprehension test items, she generated multiple *item variants* by modifying item features, including propositional density, use of passive voice, negative wording, order of information, and lexical similarity between the passage and response options – all of which were theoretically grounded in an IDM. Two hundred and seventy-eight undergraduates were given a subset of 27 items of varying types (i.e., inference, author's purpose, vocabulary in context) and the items were associated with a variety of passages (e.g., humanities, social sciences, physical sciences). Results showed that manipulation of some passage features such as increased use of negative wording significantly increased item difficulty. Others, such as altering the order of information presentation in a passage, did not significantly affect item difficulty but did affect reaction time. These results provide evidence that certain theoretically based item features directly affect cognitive processing and can be considered part of the measured construct\and, thus, should also be part of any cognitive model of the enacted construct.

In the same study, non-significant results of several manipulations challenged the validity of the cognitive model given that no direct links between theoretically relevant item features and individual item difficulty were established. Experimental manipulations such as these, when applied in item development stages, can be useful in establishing the meaning of the enacted construct measured by a test and suggest potential modifications that could strengthen the validity of score interpretations. In Gorin's (2005) study, the results were able to both confirm some cognitive components and falsify (or, at least, show no support for) other assumptions of the cognitive model of the instantiated (enacted) reading comprehension construct. Thus, experimental methods, with the explicit goal of exploring causal relationships, are uniquely positioned to test cognitive models or theories, and can allow us to seek explicitly both confirmation and falsification or refutation of different assumptions.

One practical constraint is when and how these methods can be used. In educational assessment, it is typically impractical and often unethical to administer experimentally manipulated items or other stimuli in an operational setting. But earlier in the item development process, experimentally manipulating aspects of items can pay dividends in item design decisions such as being able to select variables that have been shown to produce the best evidence possible for the target construct (e.g., Katz et al., 2000; Snow & Katz, 2010). Having a cognitive model from which to generate key

research questions and select important experimental manipulations should enhance the value of the results and the results of the experimental study should, in turn, help to refine the cognitive model of the enacted construct.

On-line Evidence: Methods and Approaches

The fields of cognitive science and neuroscience have observed developments and validations of a number of methods allowing researchers to infer cognitive processes. Many of these methods capture **protocol data**, which are sequential and dynamic time series data reflecting moment-to-moment cognition. Protocol data are quite varied – they can consist of verbal utterances made by research participants that reflect their thinking as they work through tasks, sequences of gaze patterns that reflect dynamically changing attention, or neurophysiological traces reflecting measurable neural indices that co-occur with cognitive processes. All of these data provide an observable trace over time, which can be used to infer processes such as high-level reasoning, attention, and cognitive effort. In design research and ergonomics, for example, online measures of individuals' interactions with systems have been used to infer cognitive processes and decision processes (Covey & Lovie, 1998; Ford, Schmitt, Schectman, Hults, & Hoherty, 1989). The continuous data from such methods are qualitatively different in nature from static or singular data points such as scores or response times, which can allow us to observe an *outcome* of cognition but not infer the *process* of cognition.

With this focus on process, protocol methods and data are well suited for supporting the development and validation of cognitive models. None of these methods is perfect – each can provide information about only certain aspects of cognitive processing. But in educational assessment, where the goal is increasingly not just to measure differences but to understand why individual test-takers or sub-groups differ in terms of solving a problem or answering an item, having methods that can be used to explore the process of completing a task is potentially very valuable. In the following subsections we describe some cognitive and neuroscientific methods that allow us to infer at least some aspects of that process. We attempt to highlight some strengths and limitations of these methods, and we aim to provide suggestions for how and when they may be effectively applied in assessment development.

Methods that produce verbal data. A number of research methods used in basic cognitive science as well as in other, more applied, domains involve the collection of verbal data and/or some form of self-report. Research methods in this general category include **think-aloud** or **talk-aloud methods**, **cognitive interviews**, usability studies, and, more recently borrowed from the game development world, **playtesting** (an informal small-group method that encourages naturalistic conversation among participants working together on draft tasks). There has been growing momentum for the use of think-aloud and other **verbal protocol** or self-report methods in educational assessment. Snow and Lohman (1989) were among the first to suggest their use in this field, and the approach has been subsequently echoed and implemented by others in the intervening decades. For example, at the *Educational Testing Service*, we have been using concurrent think-aloud protocols

since the 1990s to investigate validity and assessment development issues in a wide variety of domains, including logical reasoning (Enright, Tucker, & Katz, 1995), mathematics (Katz, Bennett, & Berger, 2000; Nhouyvanisvong & Katz, 1998), and architecture licensure (Katz, 1994). However, the emphasis on cognitive psychology and associated methods, including verbal protocols, presented in the *National Research Council's* 2001 report *Knowing What Students Know: The Science and Design of Educational Assessment* (Pellegrino, Chudowsky, & Glaser, 2001) brought the issue to greater visibility in the larger educational community (Leighton, 2009).

Since then, an increasing number of educational assessment researchers have noted the potential for verbal protocols to inform cognitive model building and validation as a critical step in educational assessment design (Embretson & Gorin, 2001; Erickan et al., 2010; Gorin, 2006a; Leighton, 2004, 2009; Leighton & Gierl, 2007; Mislevy, 2006). Verbal protocols applied to assessment tasks provide unique insight into individual processing, including information about student misconceptions, skill weaknesses, and uses of various problem-solving **strategies** (Leighton, 2004). In terms of the initial building and validation of a cognitive model, this approach can be particularly useful as an initial method of investigation when researchers know little about an item type.

Conversely, if a hypothesized processing model has already been developed to the degree where it can be tested, a confirmatory approach can be used to seek validation of the model. As described earlier, Erickan and her colleagues (2010) used think-aloud protocols to test hypotheses about sources of DIF across linguistically and culturally diverse groups. Explanatory DIF analysis is a form of validation of a cognitive model, in that the source of the DIF is viewed as a function of differential cognitive processing of items for different groups. In that study, the verbal protocol methods provided substantive explanations of the DIF beyond what statistical analysis revealed about the location and quantity of DIF. In addition, verbal protocol methods addressed the validity of the cognitive model underlying the test.

When planning the collection and use of verbal data for developing and validating cognitive models for educational assessment, several methodological factors are critical to consider. One important issue is whether verbalizations are being elicited *concurrently* in real-time as the task is being completed or only *retrospectively* after the task has ended. This factor is also related to assumptions about *cognitive load* (i.e., whether verbalizations produced by participants while completing a task impair performance on the task itself) and goals of the verbalizations (e.g., whether the participant is aiming to communicate clearly to the researcher or simply verbalizing his or her thoughts in a relatively unmodified form). A further factor to consider is the objectives of the research and/or the phase of task development – different methods may be appropriate depending on whether one is attempting to develop a cognitive model of the processes involved in completing a task (where the primary focus is the non-visible internal cognition of the test-taker) versus evaluating the same task for usability factors or clarity of wording (where the primary focus is the external task and its design).

Two approaches that are typically used for gathering verbal data in assessment development and research are think-aloud and cognitive interview methods. Think-aloud procedures involve a participant verbalizing his or her thoughts as they occur while working through a task. A seminal set of studies conducted in the 1970s and 1980s demonstrated empirically that, with the right methodological approach, verbalizations

could be a valid form of data for revealing cognitive processes (Ericsson & Simon, 1980, 1993). Studies have shown that certain constraints must be in place in order to ensure the validity of this kind of data. Verbalizations must be produced concurrently while doing the task, because retrospective descriptions by participants of their thinking on a previously completed task rely too much on recall and thus do not produce the same unmodified reflection of the cognitive process (Ericsson & Simon, 1993; Pressley & Afflerbach, 1995). In addition, the effort of verbalizing must be minimized so that it does not interfere with the main task, and the data accurately reflect the cognitive processes involved in task performance as normally as possible (Ericsson, 2006). Utterances made by participants should be no more than externalizations of their inner speech as they work through the task or, if the thinking is nonverbal and doesn't therefore give rise to inner speech, an unmodified verbal representation of their conscious experiences as they work through the process. Ericsson (2006) provides this example of verbalizations from a participant doing a mental arithmetic problem, which involved multiplying 36 by 24:

> OK, 36 times 24, um, 4 times 6 is 24, 4, carry the 2, 4 times 3 is 12, 14, 144, 0, 2 times 6 is 12, 2, carry the 1, 2 times 3 is 6, 7, 720, 720, 144 plus 720, so it would be 4, 6, 864.

In think-aloud methods, the interviewer or facilitator's role is simply to prompt the participant to keep talking as they are working, providing a reminder each time there is a gap in the train of verbalizations lasting more than a few seconds. Thus, facilitators should do no more than prompt test-takers to keep producing a continuous stream of verbalizations throughout the task that reflect the thoughts experienced, using non-directive reminders such as "please keep talking." The types of verbalizations that are expected are modeled for the participant prior to starting the think-aloud interview with examples provided to clearly demonstrate that verbal utterances can be broken and incomplete. Thus, verbalizations should sound similar to the kinds of inner speech that we experience internally when we are contemplating some problem, rather than the more fully formed and pre-planned external speech that we produce in order to communicate with others. The goals of this approach are to avoid inadvertent contamination of the verbal report (protocol) and by extension the cognitive process by the facilitator leading the participant with questioning.

Think-aloud methodology generates rich data and is an analysis-intensive approach, since the raw verbalizations require extensive qualitative interpretation. In the cognitive psychology literature these studies are often conducted with very small sample sizes, even as few as one individual if that individual is of special interest (for some examples, see Ericsson, 2006, pp. 236–237). In the assessment field the goal is generally to develop models that can elucidate performance at the group or population level and, thus, studying a single individual is not usually appropriate, but sample sizes for this methodology are still typically quite small.

In some cases, the goal may be to discover some of the range of strategies or processes that test-takers use to complete a task or item. Not all individuals take exactly the same approach, and with the advent of digitally captured process data from interactive tasks and items, characterizing different cognitive and behavioral strategies and approaches among test-takers has the potential to be informative for reporting purposes. Think-aloud

data captured on a smaller scale in the lab may help to inform and provide validity evidence for cognitive inferences from process data captured on a larger scale during digitally based assessments, especially when the tasks and items are designed with a cognitive model in mind (e.g., Katz, 1994; Keehner & Smith, 2013).

Cognitive interviews typically involve directed questions asked during or after completion of a task or activity. Having the opportunity to use *verbal probes* makes for a methodology that can be more targeted and directed to meet specific goals compared to think-aloud methods (Willis, 2005). This methodology is often used in the development of self-report items such as survey questions (Boeije & Willis, 2015). Research comparing the two approaches indicates that verbal probing can be a more effective method for identifying problems with questions or items and for exploring specific components of each question or item, which makes sense given that method's **affordances** for targeting issues of interest. However, compared to think-aloud protocols, cognitive interviews are less revealing of the cognition that occurs while participants are thinking about and answering items (Priede & Farrall, 2011). In practice, especially in pre-testing of survey items, verbal probing, either concurrent or retrospective, is often combined with think-aloud methods in a hybrid approach (Boeije & Willis, 2015). For example, a student may listen to or read a question and then think aloud to verbalize their reaction to it, during or after which the facilitator may interject with targeted questions. However, it should be noted that adding verbal probing to think-aloud methodology inevitably contaminates the verbal protocol data and the think-aloud data (Leighton, in press) so the trade-off should be weighed carefully. As in any research endeavor, the precise details of the methodology should depend on the research questions being asked, the type of evidence sought, and the implications and intended uses of the findings (Beatty & Willis, 2007); for a helpful discussion of when to use each method to support validity claims see Leighton (in press).

They key to eliciting verbal data about cognition is to ensure that the right kinds of prompts from the experimenter or facilitator elicit the right kinds of verbalizations from the individual participant (Ericsson & Simon, 1993; 1998). Leighton (2011) examined the **reliability** and accuracy of verbal protocol data for educational assessments and, consistent with the findings in the more historical cognitive psychology literature, concluded that the accuracy of the data is affected by several factors. Most notably, test-takers' verbalizations can be significantly influenced by the difficulty level of the items relative to the ability level of the student in addition to characteristics of the interviewer and the perceived "expertise" of that person. Again, this is a lesson about the care that must be taken in generating verbal protocol data and its interpretation when we are seeking data that accurately reflect the processes of cognition for developing and validating a model of those processes (Leighton & Gierl, 2007).

Log files. With interactive technology-based assessments, we can capture another form of process or protocol data – test-taker interactions with computer-presented tasks and items. Like verbalizations, the stream of test-taker interactive behaviors that occur during an interactive task or item (e.g., mouse clicks or taps/swipes on interactive elements, scrolling and other navigational behaviors, keystrokes, deletions, highlighting, edits, and the pauses between events) can be viewed as a kind of observable trace that reflects at least some aspects of the student's cognition as it unfolds in time

(Baker & Yacef, 2009; Mislevy, Behrens, Dicerbo, & Levy, 2012; Rupp et al., 2012). Each time a test-taker clicks on an interactive tab or button, drags and drops an object onto a target, hits PLAY to watch a video or hear an audio file, selects variables to manipulate in a simulation, or simply hits SUBMIT, NEXT, or BACK, a digital record of that action can be captured along with the task section, system settings, and a timestamp, accurate to milliseconds, showing precisely when the event happened and in what context.

In the theoretical framework of *embodied* or *externalized cognition*, we might argue that test-takers are using the external interactive tools to think (Wilson, 2002). For example, an interactive simulation that can be used to manipulate variables and run experiments allows students to represent and manipulate information and make discoveries that would not have been possible without that interactive tool. In other words, cognition is no longer occurring only inside the head; the student is now doing some of their thinking with an external representation that they can manipulate akin to using a pencil and a paper to sketch out an idea that is too complex to hold in the head) (e.g., Wilson, 2002; Zhang & Norman, 1994). As a consequence, the affordances of the interactive task or item inevitably shape cognition by affording certain kinds of representations and actions, and these interactive affordances thus influence the enacted construct. A cognitive model of the enacted construct, therefore, needs to include the kinds of externally supported cognitive processes that are available with the interactive tool or item.

When incorporating insights from **log file** data to develop or validate a cognitive model, it is important to consider different classes of actions. The most easily identified evidence is responses to traditional items (e.g., clicking on a radio button to make a selection in a multiple choice question). However, such actions lead primarily to **product data** that are really outcomes of thinking, rather than **process data**, and tell us little about the process that led to the outcome (Rupp et al., 2012).

A more informative type of action is the interactive behavior in which the test-taker engages during the intervening time between formal assessment responses. Such actions may include non-scored behaviors, which may nonetheless be construct-relevant and cognitively meaningful, and interactive behaviors that are scored according to a performance-based scoring rubric. This is currently the case in some assessments of performance-based constructs such as science inquiry practices (Gobert & Koedinger, 2011; Gobert, Sao Pedro, Baker, Toto, & Montalvo, 2012), computer programming and troubleshooting (Behrens, Mislevy, Bauer, Williamson, & Levy, 2004), digital information literacy (Katz, 2007), or technology and engineering literacy (Keehner & Smith, 2013). Log files from interactive tasks in these assessments include rich information about the decisions students make as they solve problems. For example, log files can indicate the range of sampling that students do when manipulating variables to run experiments or the order in which they take steps to fix a problem. In such cases, the behavior pattern can be considered relative to the hypothesized strategies and processes that the tasks are intended to measure. Specifically, if we see log files from students indicating use of strategies or processes other than those specified in our cognitive model of the task, we might reconsider whether our intended construct and enacted construct are in fact well aligned. We can use this information to revise our tasks in ways to align better with the intended construct by identifying and revising task elements that elicit unintended cognitive processes (Mislevy et al., 2012).

While data about these kinds of interactive behaviors have the potential to be extremely valuable in terms of validating our cognitive models, they could have even more significant implications for scoring and score reporting. Beyond validating cognitive models, we expect to see increasing use of these kinds of observations for formative purposes and even for summative purposes as part of descriptive or qualitative reporting. These data might allow actionable conclusions to be generated such as identifying the range of strategies that successful students used to reach the right answer or the kinds of errors or missteps that lower performing students exhibited on the way to an incorrect final decision. As these kinds of data start to be incorporated into reporting outputs, there may also be an increasing role for cognitive models, for translating data into meaningful and cognitively grounded reporting claims, even in large scale operational assessments (Mislevy et al., 2012).

However, analyzing log files is not simple. As Kerr, Chung, and Iseli (2011) note, the information they provide can be hard to interpret, especially in complex interactive tasks or games in which many factors are at play at any given point. Further, the task of identifying what is meaningful and what is noise is quite challenging and can sometimes only be done post hoc after all analyses are complete because action logs typically include a record of all behaviors. Log files are also typically very large (Kerr et al., 2011), although with increases in processing capacity this is not necessarily a major issue, and, moreover, the flip side of this characteristic is that more data can support better models (DiCerbo & Behrens, 2014).

Educational data mining techniques have proven helpful for log file analysis, since they can be used to identify frequently occurring patterns in large and complex data sets (Baker & Yacef, 2009; Mislevy et al., 2012; Rupp et al., 2012). Taken to the extreme, exploratory data mining could in principle require no a priori assumptions or theory about what kinds of actions might be meaningful or what they might tell us about cognition. However, many authors recommend a more balanced approach using both top-down theoretical assumptions and bottom-up data-driven discovery in combination (e.g., see Mislevy et al., 2012).

For example, having an **evidence-centered design** framework (Mislevy, Almond, & Lukas, 2003) made it possible for Kerr and Chung (2012) to identify key performance characteristics in log file data from games and **simulations** using cluster analysis, and it allowed them to interpret the patterns they found in meaningful ways that were related to the constructs of interest. But there was still room for some exploration and discovery – the fuzzy clustering approach they used also facilitated the identification of interaction patterns that were not predicted by the evidence-centered design framework. This combination of top down assumptions and bottom-up exploration, sometimes in conjunction with cognitive science methodologies such as think-aloud studies, has also proven helpful in our own experiences with log file data, in terms of inferring the cognition from the captured interactions (Keehner, Agard, Berger, Bertling, & Shu, 2014; Komsky et al., 2015; Oranje, Keehner, Mazzeo, Xu, & Kulick, 2012).

Eye-tracking. **Eye-tracking** is a technology that measures the direction of one's line of gaze with millisecond timing and millimeter accuracy, using an infrared video camera positioned a few feet away or a pair of glasses with sensors attached (Duchowski, 2003). Like think-aloud verbalizations and streams of human-computer interactions, gaze

fixations and gaze sequences are indicators of cognition as it occurs over time. Fixations and gaze sequences reflect how test-takers attend to visual information as they are processing it during the course of a task or item. But unlike think-aloud methods, eye tracking doesn't require the participant to think aloud (or direct any overt attention to their own cognitive processes), and unlike log file data, it does not depend on having a stimulus that elicits interactive behaviors.

These characteristics make eye tracking an especially valuable methodology for examining cognitive processes in reading. In a reading comprehension assessment there may be relatively little variability in available ways of interacting with the task, since reading long passages does not involve many discrete actions other than page turning or scrolling and actions associated with responding to questions. In addition, in a reading comprehension assessment there may be long pauses between page turns or other actions, thus limiting the use of log file data for making cognitive inferences. Furthermore, although thinking aloud while reading can be done under the right conditions (Pressley & Afflerbach, 1995), it can be challenging, due to factors such as competition for overlapping cognitive resources from simultaneous speech production and text processing (Baddeley, 1992) and the largely automatic nature of the process, which can make the cognition hard to verbalize (Ericsson & Simon, 1993). Eye movements are, therefore, a valuable source of evidence about the otherwise "invisible" cognitive processes of reading.

The value of eye-tracking is increasingly recognized in the educational assessment community (Gorin, 2006b; Gorin & Embretson, 2012; Mislevy, Bejar, Bennett, Haertel, & Winters, 2010; Svetina, Gorin, & Tatsuoka, 2011). Gorin (2006b) was among the first to report a preliminary eye-tracking study specifically designed for understanding assessment design. Since that time, eye-tracking has been used to study cognitive processes in a standardized science test with pre-service science teachers (Tai, Loehr, & Brigham, 2006), an elementary-level reading assessment involving text and graphics (Solheim & Uppstad, 2011), university-level image-based multiple choice problems (Tsai, Hou, Lai, Liu, & Yang, 2012), and a seventh-grade high stakes reading comprehension assessment (Knight & Horsley, 2014). It is seen as a promising technology to help us understand and differentiate student cognition in assessment items. For example, experts in a domain (e.g., high performing students in mathematics) may optimize the ways in which they scan or monitor elements of a task during problem-solving, whereas someone with lower competency may show a disorganized eye movement pattern indicative of poor knowledge and skills (Lauwereyns & d'Ydewalle, 1996; Salvucci & Anderson, 2001). Nevertheless, empirical eye-tracking studies in the context of assessment research are still rare compared to those in basic studies of reading (Rayner, 1998), mathematics (Hegarty, Mayer, & Green, 1992; Salvucci & Anderson, 2001), and other fields (Feng, 2011).

In one assessment study, Feng and colleagues (Feng et al., 2012) set out to test whether multiple choice questions induce a piecemeal reading strategy, whereby students read just enough to answer the question. They tracked the gazes of 30 university students taking a standardized reading comprehension assessment. Figure 4.2 illustrates the sequential *scanpath* of one student in the multiple choice-only testing condition; circles indicate fixation locations, with numbers indicating their temporal sequence. The student began by reading the question stem (fixation #2–9) and option 1 (fixation #10–15)

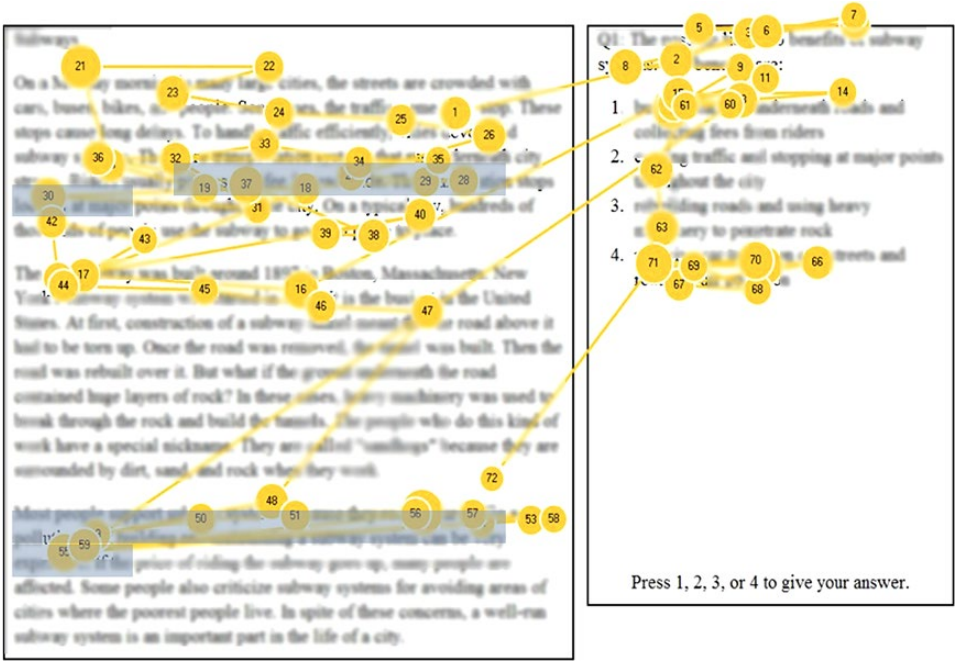

Figure 4.2 Scanpath of one reader in the multiple choice-only condition of Feng et al. (2012) applying a piecemeal reading strategy to answer the first multiple choice question. Circles indicate fixation locations, with numbers indicating their temporal sequence. Information necessary to answer the question is marked in gray.

before jumping to the second paragraph (fixation #16–17) and then the first paragraph. In lines 3 and 4 of the first paragraph the reader briefly encountered the first piece of evidence to answer the multiple choice question. The reader continued to read paragraph 1, skimmed the first sentence of paragraph 2, and then arrived at the first sentence of paragraph 3 where the other piece of the answer could be found. With that information the student went on to answer the multiple choice question. The eye movement pattern suggests that the student probably read to answer the question without having a good understanding of paragraphs 2 and 3. Feng et al. (2012) showed that asking students to write a brief summary of the passage encourages thorough reading and more coherent comprehension. For example, those who first wrote a summary of the passage spent significantly less time re-reading the text when answering the same multiple choice questions and, when they looked back at the passage, they spent less time searching in the text suggesting they knew where to find the information.

In another assessment study (Feng, Sands, Redman, Deane, & Sabatini, 2013), eighth-grade students were asked to evaluate whether a summary contained the main idea of an article they had read. The task required students to read and identify the main idea paraphrased in the summary. Figure 4.3a shows a *heatmap* that combines the eye gazes for all students who answered the question correctly. In contrast to the sequential information shown in a scanpath visualization, a heatmap shows overall dwell times in different regions of the stimulus. The region containing the main idea, as marked in the figure, received a concentration of attention

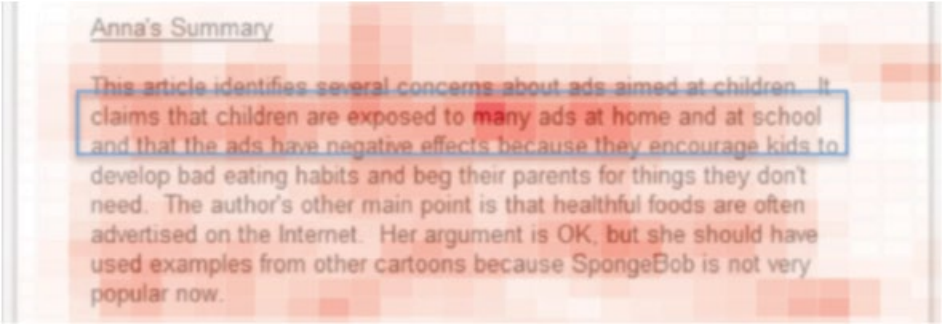

Figure 4.3a Eye gaze heatmap for students who correctly identified the main idea in the summary (Feng et al., 2013).

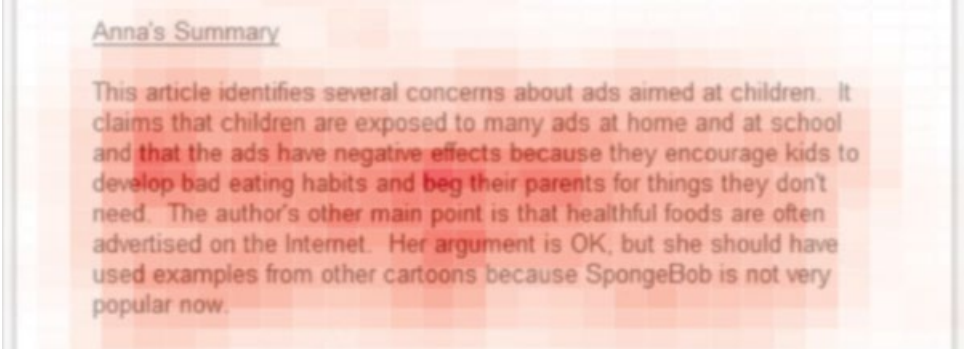

Figure 4.3b Eye gaze heatmap for students who failed to identify the main idea in the summary (Feng et al., 2013).

from students who answered correctly. In contrast, the group heatmap for students who answered the question incorrectly (Figure 4.3b) looks quite different and suggests that these students spent more time reading parts of the summary that were irrelevant to the question. Note that this is a qualitative comparison but formal statistical methods such as calculating the probability mass of gaze locations in the areas of the page where key information is present could be applied to test whether readers who identified the main idea were significantly more likely to concentrate on those areas.

We provided two examples here to illustrate the potential value of eye movement data for validating cognitive models of reading tasks. By themselves, such visualizations cannot prove or disprove a cognitive model since they are descriptive but there are also inferential statistical methods for analyzing eye movement data and making inferences about underlying cognitive processes (see e.g., Feng, 2006; Holmqvist et al., 2011). In addition, both eye-tracking and action log analyses often involve modeling the time course of behaviors, an area of assessment research undergoing rapid development.

As with log file sequences, if distinct behavioral patterns are discerned in eye movement data using statistical methods, appropriate psychological interpretations need to be assigned by the researcher. As was mentioned previously, an approach that combines exploratory and confirmatory methods may lead to the most accurate psychological interpretations (Mislevy et al., 2012).

Integrating methods. Occasionally it is possible, even in an operational context, to combine more than one of these approaches to triangulate sources of online evidence. At the *Educational Testing Service*, we had the opportunity to apply multiple complementary methods during sequential development phases of a large-scale assessment program and here we share what we learned from that experience. The assessment to be developed was focused on newly conceptualized engineering and design-related competencies at the middle school level and included extended interactive scenario-based tasks. Through successive phases of task development, we applied a range of research methodologies. Some of these were typical pre-testing activities used primarily to inform task design and refinement (e.g., playtesting, usability studies, and task tryouts). Others were separate lab-based studies designed to gather richer data about cognition, including eye movements and verbalizations gathered during or immediately after task completion. These additional cognitive studies were limited by the typical budgeting and scheduling constraints of an operational research program and therefore lacked the strict empirical rigor of a dedicated research stream such as blinded or multiple raters for inter-rater reliability estimates. Nevertheless, we believe that they provided valuable additional data for making inferences about the cognitive processes occurring during the tasks.

During the larger-scale task tryouts, field trials, and the eventual administration of this assessment, increasingly large sets of process data were captured in the form of log files of student interactions. These contained task-relevant interactions coded, contextualized by task section, and time-stamped. We ran iterative analyses on each of these data sets as they became available, using mainly descriptive statistics and methods shaped by both top-down hypotheses and informed explorations of the data. Our interpretation of the large-scale log files was strongly informed by the eye movements and verbalizations captured in the small-scale laboratory studies.

From the smaller earlier log file data sets, we created visualizations of students' sequences of actions to examine commonalities and differences in the range of observed behaviors. For the larger log file data sets, we performed *cluster analyses* to identify groups of students with similar behavioral profiles at specific points in the tasks, and again we visualized the sequences to help make inferences about the differing cognitive processes that were reflected in each group. The actions (i.e., log file events) that we entered into the cluster analyses were informed by informal cognitive task analyses, which generated hypotheses about potentially meaningful construct-relevant behaviors at specific places in the tasks. But they were also strongly influenced by the data we collected in the cognitive labs, which helped to corroborate or suggest the likely cognitive meaning of these behaviors when we observed them in conjunction with students' gaze patterns and their verbal descriptions of their thinking.

Several times we observed some pattern in the log file data that appeared to be analogous to something we had seen students doing, or heard them describing, in the cognitive laboratories. Similarly, the eye movement data helped us to infer what might be going on during the apparent gaps in the action sequences when the log file data suggested that students weren't doing anything for some period of time. In fact, the eye movements from the cognitive studies indicated that particular pauses at certain points in the task were often filled with some meaningful and construct-relevant behavior such as inspecting data and deciding on the next trial to run, but this would not have been visible to us except via these methods. Having the cognitive lab data thus allowed us to connect some of the "disembodied" behaviors in the log files to the "embodied" cognitive data we had observed in person with students in these studies, allowing us to make inferences about behavior patterns that might otherwise have been difficult to account for. Table 4.1 shows the research phases, the methodologies and goals of the studies, and the kinds of data that were available at different stages during assessment development and administration.

We believe that the insights we gained from adding these cognitive methods to our typical operational research paradigms and having had the opportunity to combine those insights with iterative rounds of log file analyses enabled us to more confidently make inferences about what students were doing and how they were thinking during these complex, rich, and interactive tasks. For example, in one task, which involved students freely running trials and manipulating variables to gather data in order to reach a conclusion related to how a system works, we were able to identify clusters of students with distinct behavioral profiles. However, although the log files of the behaviors distinguished these subgroups, the question of what the behaviors meant and why these individuals were behaving in different ways in terms of the underlying cognition would have been difficult to answer with only the disembodied log files of interactions. But when we put these clusters together with the problem-solving processes that we had observed in the cognitive lab, especially with the process tracing data from eye movements and students' verbalizations about their own thinking, we were able to more confidently make an inferential leap and propose possible cognitive processes – and in some cases non-cognitive factors – that may have contributed to the behaviors we observed.

Of course, we were not directly observing cognition with any of these methods. But when we have multiple sources of online data that provide a trace of the process (e.g., eye movements, verbalizations, log files of interactions), the inferences we can draw can move us closer to being able to theorize about some aspects of student cognition in these kinds of tasks. We would argue that the general approach of triangulating different sources of online evidence about cognition is valuable for any assessment seeking to report out cognitively meaningful conclusions from log file data. To this end, it is important to gather data, as resources allow, across the multiple phases of assessment development, using cognitive methods and, if possible, incorporating theoretically driven experimental manipulations of task components (which we were not able to do in this instance, but which has been done in some assessment development efforts) (e.g., Katz et al., 2000; Snow & Katz, 2010).

Table 4.1 Research phases in a large-scale operational assessment program.

Study and sample size per task presented	*Playtesting studies (N = 10)*	*Usability studies (N = 10)*	*Small-scale tryouts (N = 20)*	*Large-scale tryouts (N = 250)*	*Cognitive studies (N = 9)*	*Field pilot testing (N = 1300)*	*Final operational assessment*
Purpose	Observe student perceptions of, and interactions with, tasks; identify design issues	Test student interactions with task, interface, device; identify usability issues	First look at interactive behaviors; begin to identify patterns and develop theory	First look at log files; identify behavioral markers; analyze behaviors, expand theory	Gather evidence for cognitive process; refine theory; account for observed behaviors	Confirm relationship between scores and log file data; decide extended reporting targets	Gather data for reporting out-score reporting or other descriptive or qualitative reporting
Method	Ethnographic/ observational with some probing; small groups	Structured behavioral protocol plus cognitive interviewing	Uninterrupted task completion, unobtrusive screen capture	Uninterrupted task completion, log file action capture	Eye tracking in task, post hoc verbal report of process with relay of gaze via cursor	Uninterrupted task completion, log file action capture	Uninterrupted task completion, log file action capture
Types of data	Spontaneous undirected talk among users; verbal responses to ad hoc probing	Interactions with task and system; answers to targeted scripted verbal questions	Naturalistic unmodified on-task behavior (screen capture video) and item responses	Log files of user interactive behaviors; item responses and raw scores	Eye movements overlaid on task capture; verbal report data	Raw scores, item responses, log files of interactions. Student survey responses	Item responses, log files of interactions. Student survey responses, scale scores

Conclusion

The goal of this chapter was to show the value of cognitive models in assessment and describe ways in which it is feasible to gather a range of evidence to develop and validate models of student performance. The research studies and methodologies described in this chapter demonstrate that this is an attainable goal and many researchers and professionals in the field are already building empirical data-gathering into their work. Like Mislevy (2006), we believe the role of cognitive models in assessment will continue to grow for use in item design, scoring, and reporting, as well as for the interpretation of new evidence types such as log file data. Therefore, this meshing of the goals and methods of cognitive science and assessment science should become increasingly common and increasingly integral to assessment efforts of all types. By building a repertoire of methods for gathering cognitively informative data, applying these wherever feasible, and doing the necessary work to draw out inferences about cognition, we can gain important information for improving our assessments and their conclusions. As a result, we will incrementally help to build a new body of knowledge that will contribute to the fields of assessment, education, and cognitive science.

References

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, *22*(3), 261–295.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly: An International Journal*, *2*(1), 1–34.

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556–559.

Baker, R. S. J. D., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, *1*(1), 3–17.

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly*, *71*(2), 287–311.

Behrens, J. T., Mislevy, R. J., Bauer, M., Williamson, D. M., & Levy, R. (2004). Introduction to evidence centered design and lessons learned from its application in a global e-learning program. *International Journal of Testing*, *4*(4), 295–301.

Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2002). *A feasibility study of in-the-fly item generation in adaptive testing* (GRE Board Professional Rep. No. 98-12P). Princeton, NJ: ETS.

Boeije, H., & Willis, G. (2015). The Cognitive Interviewing Reporting Framework (CIRF). *Methodology*, *9*(3):87–95. doi: 10.1027/1614-2241/a000075

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing*, *15*(2), 119–157.

Buck, G., Tatsuoka, K., & Kostin, I. (1997). The subskills of reading: Rule-space analysis of a multiple choice test of second language reading comprehension. *Language Learning*, *47*(3), 423–466.

Busemeyer, J. R., & Diederich, A. (2010). *Cognitive modeling*. Los Angeles, CA: Sage.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and ant validation by the Multitrait-multimethod Matrix. *Psychological Bulletin*, *56*(2), 81–105.

Covey, J. A., & Lovie, A. D. (1998). Information selection and utilization in hypothesis testing: A comparison of process-tracing and structural analysis techniques. *Organisational Behavior and Human Decision Processes*, *75*, 56–74.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302.

DiCerbo, K. E., & Behrens, J. T. (2014). *Impacts of the digital ocean on education*. London: Pearson.

Duchowski, A. T. (2003). *Eye tracking methodology: Theory and practice*. New York, NY: Springer.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*(1), 179–197.

Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika*, *49*(2), 175–186.

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

Embretson, S. E. (2010). *Measuring psychological constructs: Advances in model-based approaches*. American Psychological Association.

Embretson, S. E., & Gorin, J. S. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, *38(4),* 343–368.

Enright, M. E., Tucker, C., & Katz, I. R. (1995). *A cognitive analysis of solutions for verbal, informal, and formal-deductive reasoning problems* (ETS Rep. No. RR-95-6). Princeton, NJ: Educational Testing Service.

Enright, M. K., Morely, M., & Sheehan, K. M. (2002). Items by design: The impact of systematic variation on item statistical characteristics. *Applied Measurement in Education*, *15*(1), 49–74.

Erickan, K., Arim, R., Law, D., Domene, J., Gagnon, F., & Lacroix, S. (2010). Application of think aloud protocols for examining and confirming sources of differential item functioning identified by expert reviews. *Educational Measurement: Issues and Practice*, *29*(2), 24–35.

Ericsson, K. A. (2006). Protocol analysis and expert thought: Concurrent verbalizations of thinking during experts' performance on representative tasks. In K. A. Ericsson, N. Charness, P. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performanc*. (pp. 223–241). Cambridge, UK: Cambridge University Press.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*(3), 215–251.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (rev. ed.). Cambridge, MA: MIT Press.

Ericsson, K. A., & Simon, H. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity*, *5*(3), 178–186.

Feng, G. (2006). Reading eye movements as time-series random variables: A stochastic model. *Cognitive Systems Research*, *7*(1), 70–95.

Feng, G. (2011). Eye-tracking: A practical guide for developmental researchers. *Journal of Cognition and Development*, *12*(1), 1–12.

Feng, G., Gorin, J., Sabatini, J., O'Reilly, T., Walls, C., & Bruce, K. (2012). *Reading for understanding: How comprehension facilitates answering questions, and what questions enhance understanding*. Presentation at Annual Meeting of the Society for Scientific Study of Reading, Montreal, Canada.

Feng, G., Sands, A. Redman, M. Deane, P., & Sabatini, J. (2013, July). *Understanding innovative reading assessments through eye-tracking and verbal reports*. Presentation at the Annual Scientific Study of Reading meeting, Hong Kong.

Ferrara, S., Duncan, T. G., Freed, R., Vélez-Paschke, A., McGivern, J., Mushlin, S., … & Westphalen, K. (2004). *Examining test score validity by examining item construct validity: Preliminary analysis of evidence of the alignment of targeted and observed content, skills, and cognitive processes in a middle school science assessment*. Paper presented at the 2004 Annual Meeting of the American Educational Research Association.

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

Ford, J., Schmitt, N., Schectman, S., Hults, B., & Hoherty, M. (1989). Process tracing methods: Contributions, problems, and neglected research questions. *Organisational Behavior and Human Decision Processes*, *43*, 75–117.

Frederiksen, N. (1986). Construct validity and construct similarity: Methods for use in test development and test validation. *Multivariate Behavioral Research*, *21*(1), 3–28.

Gobert, J., & Koedinger, K. (2011). *Using model-tracing to conduct performance assessment of students' inquiry skills within a Microworld*. Paper presented at the *Society for Research on Educational Effectiveness*, Washington, D.C., September 8-10.

Gobert, J. D., Sao Pedro, M. A., Baker, R. S. J. D., Toto, E., & Montalvo, O. (2012). Leveraging educational data mining for real time performance assessment of scientific inquiry skills within microworlds. *Journal of Educational Data Mining*, *4*(1), 111–143.

Gorin, J. S. (2005). Manipulation of processing difficulty on reading comprehension test questions: The feasibility of verbal item generation. *Journal of Educational Measurement*, *42*, 351–373.

Gorin, J. S. (2006a). Item design with cognition in mind. *Educational Measurement: Issues and Practice*, *25*(4), 21–35.

Gorin, J. S. (2006b). *Using alternative data sources to inform item difficulty modeling*. Paper presented at the 2006 Annual Meeting of the National Council on Educational Measurement.

Gorin, J. S., & Embretson, S. E. (2006) Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement*, *30*(5), 394–411.

Gorin, J. S., & Embretson, S. E. (2012). Using cognitive psychology to generate items and predict item characteristics. In M. Gierl and T. Haladayna (Eds.), *Automatic item generation: Theory and practice* (pp. 136–156). London: Routledge/Taylor and Francis Group.

Hartz, S. M. (2002). A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality. *Dissertation Abstracts International: Section B: The Sciences and Engineering*, *63*(2-B), 864.

Hegarty, M., Mayer, R. E., & Green, C. E. (1992). Comprehension of arithmetic word problems: Evidence from students' eye fixations. *Journal of Educational Psychology*, *84*(1), 76–84.

Holmqvist, K., Nystrom, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. New York, NY: Oxford University Press.

Kane, M. T. (2008). Terminology, emphasis, and utility in validation. *Educational Researcher*, *37*(2), 76–82.

Katz, I. R. (1994). Coping with the complexity of design: Avoiding conflicts and prioritizing constraints. In A. Ram and K. Eiselt (Eds.), *Proceedings of the sixteenth annual conference of the Cognitive Science Society* (pp. 485–489). Hillsdale, NJ: Lawrence Erlbaum Associates.

Katz, I. R. (2007). Testing information literacy in digital environments: ETS's iSkills Assessment. *Information Technology and Libraries*, *26*(3), 3–12.

Katz, I. R., Bennett, R. E., & Berger, A. (2000). Effects of response format on difficulty of SAT-Mathematics items: It's not the strategy. *Journal of Educational Measurement*, *37*(1), 39–57.

Katz, I. R., Lipps, A., & Trafton, J. G. (2002). *Factors affecting difficulty in the generating examples item type* (ETS Rep. No. RR-02-07). Princeton, NJ: Educational Testing Service.

Katz, S. & Lautenschlager, G. J. (1994). Answering reading comprehension items without passages in the SAT-I, the ACT, and the GRE. *Educational Assessment*, *2*, 295–308.

Katz, S. & Lautenschlager, G. J. (2001). The contribution of passage and no-passage factors to item performance an on the SAT reading task. *Educational Assessment*, *7*, 165–176.

Katz, I. R., Martinez, M. E., Sheehan, K., & Tatsuoka, K. K. (1998). Extending the rule space methodology to a semantically-rich domain: Diagnostic assessment in architecture. *Journal of Educational and Behavioral Statistics*, *23*(3), 254–278.

Keehner, M., Agard, C., Berger, M., Bertling, J., & Shu, Z. (2014). *Analyzing interactivity, performance, and background data from the NAEP TEL Wells task*. Federal Research Memorandum on NAEP Task Component, Institute of Education Sciences (IES) of the US Department of Education, Contract Award No. ED-IES-13-C-0015.

Keehner, M., & Smith, L. (2013). *Connecting actions, cognitions, and measurement: The role of cognitive science in NAEP TEL task development*. Paper presented at the Annual Meeting of the National Council on Measurement in Education. San Francisco, CA.

Kerr, D., & Chung, G. K. (2012). Identifying key features of student performance in educational video games and simulations through cluster analysis. *JEDM-Journal of Educational Data Mining*, *4*(1), 144–182.

Kerr, D., Chung, G. K. W. K., & Iseli, M. R. (2011). *The feasibility of using cluster analysis to examine log data from educational video games* (CRESST Report 791). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Kintsch, W., & vanDijk, A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*, 363–394.

Knight, B. A., & Horsley, M. (2014). A new approach to cognitive metrics: Analysing the visual mechanics of comprehension using eye-tracking data in student completion of high-stakes testing evaluation. In M. Horsley (Ed.), *Current trends in eye tracking research* (pp. 287–296). New York, NY: Springer International Publishing.

Komsky, J., Kerr, D. Keehner, M., Cayton-Hodges, G. A., Katz, I. R., Koster van Groos, J., & Brockway, D. (2015, August). *Exploring the use of interactive science simulations for assessment*. Poster presented at the CRESST Conference August 2015: Making Games and Technology Work for Learning, Assessment and Instruction, Redondo Beach, CA.

Lauwereyns, J., & d' Ydewalle, G. (1996). Knowledge acquisition in poetry criticism: The expert's eye movements as an information tool. *International Journal of Human-Computer Studies*, *45*(1), 1–18.

Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, *23*(4), 6–15.

Leighton, J. P. (2009, April). *How to build a cognitive model for educational assessments*. Paper presented at the 2009 Annual Meeting of the National Council on Measurement in Education. San Diego, CA.

Leighton, J. P. (2011, April). *Item difficulty and interviewer knowledge effects on the accuracy and consistency of examinee response process in verbal reports*. Paper presented at the 2011 Annual Meeting of the American Educational Research Association. New Orleans, LA.

Leighton, J. P. (in press). Collecting, analyzing and interpreting verbal response process data. In K. Ercikan and J. Pellegrino (Eds.), National Council on Measurement in Education (NCME) Book Series - *Validation of Score Meaning in the Next Generation of Assessments*. London: Routledge.

Leighton, J. P., & Gierl, M.J. (2007). Verbal reports as data for cognitive diagnostic assessment. In J. P. Leighton and M. J. Gierl (Eds.) *Cognitive diagnostics assessment for education: Theory and applications* (pp. 146–172). New York, NY: Cambridge University Press.

Markman, A. B. (1999). *Knowledge representation*. Mahwah: NJ: Lawrence Erlbaum Associates.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, *30*, 955–966.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11.

Miller, K. S. (2000). Academic writers on-line: Investigating pausing in the production of text. *Language Teaching Research*, *4*(2), 123–148.

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. *Educational Measurement*, *4*, 257–305.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series, 2003*(1), i–29.

Mislevy, R. J., Behrens, J. T., Dicerbo, K. E., & Levy, R. (2012). Design and discovery in educational assessment: Evidence centered design, psychometrics, and data mining. *Journal of Educational Data Mining*, *4*(1), 11–48.

Mislevy, R. J., Bejar, I. I., Bennett, R. E., Haertel, G. D., & Winters, F. I. (2010). Technology supports for assessment design. *International Encyclopedia of Education*, *3*, 56–65.

Nhouyvanisvong, A., and Katz, I. R. (1998). The structure of generate-and-test in algebra problem solving. *Proceedings of the Twentieth Annual Meeting of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Oranje, A., Keehner, M., Mazzeo,J., Xu, X., & Kulick, E. (2012). *An adaptive approach to group-score assessments*. Federal Research Report, Task Order Component, IES contract ED-07-CO-0107.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing What Students Know. The Science and Design of Educational Assessment*. Washington, DC: National Academy Press.

Powers, D. E., & Wilson, S. T. (1993). *Passage dependence of the New SAT reading comprehension questions* (College Board Report No. 93-3). New York, NY: College Board.

Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading*. Hillsdale, NJ: Erlbaum.

Priede, C., & Farrall, S. (2011). Comparing results from different styles of cognitive interviewing: "Verbal Probing" vs. "Thinking Aloud." *International Journal of Social Research Methodology*, *14*(4), 271–287.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*, 372–422.

Rupp, A. A., Levy, R. Dicerbo, K. E., Sweet, S. J., Crawford, A. V., Calico, T., … Behrens, J. T. (2012). Putting ECD into practice: The interplay of theory and data in evidence models within a digital learning environment. *Journal of Educational Data Mining*, *4*(1), 49–110.

Salvucci, D. D., & Anderson, J. R. (2001). Automated eye-movement protocol analysis. *Human-Computer Interaction*, *16*(1), 39–86. doi:10.1207/s15327051hci1601_2

Sheehan, K. M., & Ginther, A. (2001). *What do passage-based multiple choice verbal reasoning items really measure? An analysis of the cognitive skills underlying performance on TOEFL reading comprehension items*. Paper presented at the 2001 Annual Meeting of the National Council on Educational Measurement. Seattle, WA.

Snow, E., & Katz, I. (2010). Using cognitive interviews and student response processes to validate an interpretive argument for the ETS iSkills™ assessment. *Communications in Information Literacy*, *3*(2), 99–127.

Snow, R. E., & Lohman, D. F. (1989). *Implications of cognitive psychology for educational measurement*. Stanford, CA: Center for Educational Research at Stanford.

Solheim, O. J., & Uppstad, P. H. (2011). Eye-tracking as a tool in process-oriented reading test validation. *International Electronic Journal of Elementary Education*, *4*, 153–168.

Svetina, D., Gorin, J. S., & Tatsuoka, K. K. (2011). Defining and comparing the reading comprehension construct: A cognitive-psychometric modeling approach. *International Journal of Testing*, *11*(1), 1–23.

Tai, R. H., Loehr, J. F., and Brigham, F. J. (2006). An exploration of the use of eye, gaze tracking to study problem, solving on standardized science assessments. *International Journal of Research and Method in Education*, *29*(2), 185–208.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational Statistics*, *10(1)*, 55–73.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, R. L. Brennan (Eds.) *Cognitively diagnostic assessment* (pp. 327–359). Hillsdale, NJ: Lawrence Erlbaum Associates.

Tatsuoka, K. K., Corter, J. E., & Guerrero, A. (2004). *Coding manual for identifying involvement of content, skill, and process subskills for the TIMSS-R 8th grade and 12th grade general mathematics test items*. Technical Report. New York, NY: Department of Human Development, Teachers College, Columbia University.

Tsai, M. J., Hou, H. T., Lai, M. L., Liu, W. Y., & Yang, F. Y. (2012). Visual attention for solving multiple-choice science problem: An eye-tracking analysis. *Computers and Education*, *58*(1), 375–385.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS Research Report No. RR-05-16). Princeton, NJ: Educational Testing Service.

von Davier, M. (2011). *Equivalency of the DINA model and a constrained general diagnostic model*. Research Report 11-37). Princeton, NJ: Educational Testing Service. Retrieved from http://www. ets. org/Media/Research/pdf/RR-11-37. pdf

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika*, *45*, 479–494.

Willis, G. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin and Review*, *9*(4), 625–636.

Wilson, M., De Boeck, P., and Carstensen, C. H. (2008). Explanatory item response models: A brief introduction. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 83–110), Toronto: Hogfre & Huber.

Xu, X., & von Davier, M (2006). *Cognitive diagnosis for NAEP proficiency data* (ETS Research Report No. RR-06-08). Princeton, NJ: Educational Testing Service.

Zhang, J., & Norman, D. A. (1994). Representations in distributed cognitive tasks. *Cognitive Science*, *18*(1), 87–122.

5

An Integrative Framework for Construct Validity

Susan Embretson

Research on cognitively-based approaches to assessment have become increasingly prevalent in the educational and psychological testing literature. Studies that relate cognitive principles for **item design** and associated response processes to assessment have appeared for a variety of **item** types (Daniel & Embretson, 2010; Gierl & Haladyna, 2013; Gorin, 2006; Goto, Kojiri, Watanabe, Iwata, & Yamada, 2010; Newstead, Brandon, Handley, Dennis, & Evans, 2006; Rijmen & DeBoeck, 2001). Understanding these principles is important for contemporary directions in measurement for a variety of purposes.

First, **item generation**, both algorithmic and automatic, is becoming an increasingly prominent method to produce large pools of items (Bejar, 2002; Embretson, 2002; Gierl, Zhou, & Alves, 2008; Luecht, 2013; Mortimer, Stroulia, & Yazdchi, 2013; Singley & Bennett, 2002). Embedding cognitive principles into the generation of item structures or **item families** and associated databases or item pools can help anticipate the psychometric properties of items.

Second, **cognitively diagnostic assessment** (e.g., Leighton & Gierl, 2007a; Rupp, Templin, & Henson, 2010) is increasingly applied in a variety of settings to assess examinee possession of skills or **attributes**. In this confirmatory approach to assessment, cognitively-grounded characterizations of items by required **knowledge, skills, and abilities**, or other kinds of cognitively-grounded personal characteristics – often called **attributes** in generic terms – are used. These attributes are used in the associated **measurement models** to characterize learners according to their level of mastery or possession of these attributes.

Third, modern **test blueprints** make increasingly more explicit references to cognitive principles for item design. In contrast, more traditional test blueprints often contain only general specifications that do not fully specify relationships between cognitive complexity of items and item content. Related to this, traditional item

The Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications, First Edition.
Edited by André A. Rupp and Jacqueline P. Leighton.