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.

development often has been considered an artistic process, depending on the insights and creativity of individual item writers. Although item writers typically are provided general guidelines for item writing, specifications for item format and style are often made more salient to them than are specifications for content features that may impact the level and sources of the cognitive complexity for items.

As a result, new items are often developed by simply instructing item writers to produce items that are "similar" to old items. This guideline certainly helps somewhat to assure that the psychometric properties for the new items are similar to the existing items for the trait as currently measured. However, this approach begs the question about the research base for knowing how item content is related to the **intended construct**. Even for aptitude tests with more detailed specifications such as the *Assembling Objects* test (Defense Manpower Data Center [DMDC], 2008), a spatial ability test in the *Armed Services Vocational Aptitude Battery* (ASVAB), a wide range of item content with a wide range of levels and sources of cognitive complexity is technically feasible for a given test form or item pool.

In contrast, contemporary achievement tests, especially tests used for high-stakes decision making, typically have more detailed test blueprints to specify item content. However, even with these more detailed item specifications, it is not necessarily clear that items written for the same category have the same levels of item difficulty or sources of cognitive complexity. Thus, in the traditional approach to developing both psychological tests and achievement tests, reviews by panels of experts and empirical tryout of items are essential to assure item and test quality. This is an expensive process that creates a bottleneck in the test development process; limited spaces on operational tests are available for tryout. Approaches to test development that could reduce this evaluation process would be desirable.

In this chapter I present an integrated and interactive framework for **construct validity** that includes all five core aspects elaborated in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) (or *Standards* for short). The proposed construct validity framework is general while other more **principled assessment** design frameworks discussed elsewhere in this *Handbook* provide more specific procedural details. My **validity** framework is interactive in that it includes several categories of interrelated aspects of validity that can impact the five core aspects of construct validity. In the first section of this chapter, I review the framework and its constituent components and illustrate the utility of this framework for assessment development and use with a few examples in my discussions. In the second section of this chapter, I apply the framework to a systematic review of evidence for the construct validity of a new form of a **fluid intelligence** test whose items were generated automatically using cognitive complexity variables grounded in previous empirical research.

The relationship of cognitive principles to the construct validity of assessments are of course also of concern considered elsewhere in this volume. These include how **cognitive model** approaches can be specified and **operationalized** to link assessments to theories of learning and cognition, and how evidence and content lead to intended interpretations. A later chapter in this *Handbook* explicates alternative methods by which cognitive complexity can be investigated in tests. It notes that understanding the sources of cognitive complexity in items supports the substantive aspect of construct validity that is related to **response processes**. The *Handbook* also compares various systems of

principled assessment that involve cognitive psychology findings to provide a general framework for understanding the potential impact of cognitively-based approaches to assessment.

Unified Framework for Construct Validity

In this section, I present a unified conceptual framework for construct validity to explicate how test development activities coupled with relevant background aspects, particularly those based on cognitive principles, can impact core aspects of construct validity. This framework is shown in Figure 5.1, which represents a reconceptualization of an earlier model (Embretson, 2007) to more clearly separate individual aspects of construct validity and to tease out those validity aspects and associated design activities that test developers can more directly manipulate.

Specifically, shown in the five circles on the right side of Figure 5.1 are the five core aspects of construct validity currently recognized in the *Standards*: *content*, *response processes, internal structure, relationships to other variables*, and *consequences*. The five aspects are furthermore organized into *internal* and *external* aspects. The internal versus external distinction is equivalent to an earlier conceptualization of **construct representation** versus **nomothetic span**, respectively (Embretson, 1983). External relationships have long been the most salient aspect of construct validity in the nomological network (Cronbach & Meehl, 1955).

In the conceptualization in Figure 5.1, however, the *relationship to other variables* and *consequences* aspects determine the significance and importance of the test as a measure of individual differences. The *external* relationships are preceded by *internal* aspects. That is, the *internal* aspects of validity, which include the *response processes*, *content*, and *internal structure* aspects, determine the meaning of the

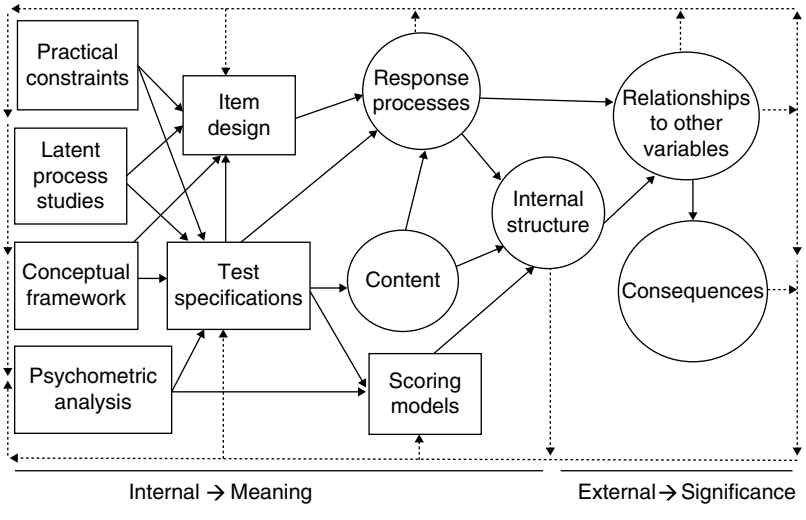

Figure 5.1 Relationship between aspects of construct validity and other variables.

trait(s) measured by the test and thus impact the significance and importance of the test as captured by *relationships to other variables* and *consequences* aspects.

Shown in the seven rectangles on the left side of Figure 5.1 are test development activities and antecedent background aspects that can impact the five core aspects of construct validity shown in the circles. Here it is useful to distinguish between four background aspects of test development processes, which are *practical constraints, latent process studies, conceptual framework*, and *psychometric analysis*; these aspects can be less directly manipulated by a test developer. In contrast, the activities of *item design*, *test specifications*, and *scoring models* can be more directly manipulated.

Core Aspects of Construct Validity

In this first sub-section, I elaborate on the five core aspects of construct validity discussed earlier in the chapter with special attention to the potential contribution of cognitive principles to best practices. The five aspects have historically had differing emphases in supporting construct validity for aptitude and achievement tests; thus, where appropriate, I elaborate on these differences. However, even though the validity framework that I present is relevant to personality and attitude measurements, I will not discuss these types of tests in this chapter in detail due to space limitations.

Content aspect. The *content* aspect of construct validity is concerned with the representation of skills, knowledge, and attributes on the test through the features of items. Relevant evidence for this aspect of validity includes the test blueprints and other test specifications as well as judgments of item appropriateness for the various specifications. As pointed out in the *Standards*, the content specifications should support the intended construct and purpose of the test.

For achievement tests, hierarchically structured blueprints are currently state-of-the-art. In mathematical achievement tests, for example, three levels are typical, with broad content categories at the top (e.g., *Number*, *Algebra*, *Geometry*, and *Data*) and more narrowly defined skill sets or competencies are the lower levels. That is, skills are the primary focus even though the cognitive complexity level of items is also typically specified, often using global categories such as those derived from Bloom's (1956) taxonomy or Webb's (1997, 2002) depth of knowledge level taxonomy.

However, while specifying cognitive complexity may seem to incorporate cognitive psychology principles, Leighton and Gierl (2007b) point out that the actual processes and skills applied by examinees to items within a category may differ substantially from the postulated specifications by experts and is subject to empirical verification. Thus, the existence of a formal blueprint alone is not sufficient evidence for the test content aspect of validity. If the sources of cognitive complexity have been studied for the item type, then the content features that determine them can be specified.

Finally, it should be noted that test administration and scoring conditions are also part of the *content* aspect and are included in the *Standards*. For example, extended versus minimal instructions, short versus longer time limits, and other such conditions, are part of the test content specifications. These variables also can impact the thought processes engaged in by examinees and hence impact the *response processes* aspect of validity.

Response processes. The *response processes* aspect of validity concerns evidence about the cognitive activities engaged in by the examinees. These cognitive activities are assumed to be essential to the meaning of the test as shown in Figure 5.1. Relevant evidence can be obtained from correlational and predictive analyses such as **item difficulty modeling** and **response time modeling** as well as more direct methods to observe individual examinees' processing such as **eye-tracking**, video analysis, and concurrent and retrospective **think-aloud** or other **verbal reports**. In Figure 5.1, the *content* aspect of validity is assumed to have a causal impact on the *response processes* aspect because examinees determine how to respond based on item and test content. However, as noted in the last sub-section, the actual processes applied by examinees' are not necessarily the theoretically intended processes. Thus, empirical evidence for the appropriateness of the applied response processes is needed.

Internal structure. The *internal structure* aspect of construct validity includes evidence for internal consistency **reliability**, test dimensionality, and **differential item/bundle/test/feature functioning**. It can be shown that empirical item properties, particularly item difficulty and **item discrimination**, directly impact the various indices of internal consistency. If these psychometric properties are in turn related to item features that impact cognitive complexity, the magnitude of the reliability indices may be impacted through item design.

Three different frameworks for estimating reliability are generally available, which include **classical test theory** (CTT) (e.g., Lord & Novick, 1968), **generalizability theory** (e.g., Brennan, 2001), and **item response theory** (IRT) (e.g., de Ayala, 2009). For many tests, any of these frameworks will be feasible. For example, the CTT approach typically involves assessing internal consistency with a statistic called *Cronbach's alpha*. Lord and Novick (1968) presented derivations to show that ρ_{alpha} can be calculated directly from item statistics as follows:

$$\rho_{\text{alpha}} = \left[k/\left(k-1\right)\right]\left[1-\left(\Sigma\sigma_i^2/\left(\Sigma\sigma_i\rho_{ix}\right)^2\right)\right] \tag{5.1}$$

where k is the number of items, σ_i^2 is item variance, and where ρ_{ix} is the point-biserial correlation of item i with total score. Thus, if the cognitive features of items are related to item difficulty and discrimination as shown in Equation 5.1, impact on internal consistency as estimated via ρ_{alpha} can be anticipated.

The generalizability theory framework, which can be viewed as an extension of the CTT framework, is concerned with identifying assessment conditions that yield a desired level of reliability. The overall index of generalizability depends on the variances associated with items and persons as well as their interactions. Because the CTT indices of internal consistency (Hoyt, 1941) are special cases of generalizability theory, cognitive complexity features can impact generalizability indices in a similar manner as for CTT.

Within the IRT framework, the item parameter estimates and the frequency of various trait estimates are used to quantify reliability. For example, for the **two-parameter logistic** (2PL) model, the probability that a person with trait level θ_j responds correctly to item i with difficulty β_i and discrimination α_i is given as follows:

$$P\left(\theta\right) = \frac{\exp\left(\alpha_i\left(\theta_j-\beta_i\right)\right)}{1+\exp\left(\alpha_i\left(\theta_j-\beta_i\right)\right)} \tag{5.2}$$

Measurement error variance for person *j* in IRT depends on the probability of passing items, *P(θ)* summed over *M* items, and, given the item parameter estimates, is computed as follows:

$$\sigma^2_{\varepsilon_j} = 1/\Sigma_{i=1}^{M}\alpha_i^2 P(\theta)\left(1-P(\theta)\right) \tag{5.3}$$

The mean error variance, $\overline{\sigma^2_{\varepsilon}}$, and the variance of estimated person scores, θ are used to compute the empirical reliability for the test as follows:

$$\rho_{tt} = \sigma^2_{\theta} / \left(\sigma^2_{\theta} + \overline{\sigma^2_{\varepsilon}}\right) \tag{5.4}$$

Given trait level and item parameter estimates, the empirical reliability can be anticipated for other combinations of items. Thus, as for the two other approaches to reliability, the cognitive complexity features of items can be used to control the empirical reliability for a test under this framework if those features have been empirically related to item difficulty and item discrimination.

As shown in Figure 5.1, both the *response processes* and *content* aspects of validity impact the *internal structure* aspect of validity. Thus, the *internal structure* aspect can be related to cognitive complexity variables. For example, specifying several independent sources of cognitive complexity can lead to heterogeneous test content that lowers item interrelationships overall and results in multidimensionality. Similarly, an insufficient range of cognitive complexity can lead to lower internal consistency by reducing item variance. Incomplete instructions to examinees may also lead to **construct-irrelevant response processes** applied to item solving and hence, in turn, leads to lower item interrelationships and lower internal consistency.

Relationship to other variables. A major external aspect of construct validity is the *relationships to other variables* aspect, which refers to the patterns of relationships of the test scores to other trait scores and empirical criteria as well as to examinee background variables that are related to the trait(s) represented by the test score(s) (e.g., demographics, prior experience, motivation measures). Empirical evidence on the relationships to other variables should be consistent with the goals of measurement to support construct validity according to the *Standards*.

The *external* aspects are clearly impacted by the *internal* aspects of construct validity. Specifically, the *internal structure* and *response processes* aspects of construct validity impact the *relationships to other variables* aspect directly while the *content* aspect impacts external relationships indirectly through the *internal structure* and *response processes* aspects. Thus, the impact of cognitive psychology principles to test development is apparent through the internal aspects of validity. If the external relationships of test scores are inconsistent with the goals of measurement, then changes in variables that impact the internal aspects are needed.

Inappropriate external correlates could provide important feedback to revise test specifications and item designs in order to create a test with more appropriate external relationships. For example, strong correlations of scores from a fluid intelligence test with scores from a vocabulary knowledge test would not support the test as purely measuring the construct of interest. Similarly, if scores from a test of contextualized

mathematical problem solving correlated too highly with scores from a test of English language ability, an analysis of the item features that impact verbal processing complexity for the mathematics items could lead to revision of test content. Items could be redesigned or additional item selection procedures could be applied to alter test content. In this example, items with higher reading levels or requiring more inferences for comprehension (e.g., selected from a pool using tools such as latent semantic analysis) should no longer be included on the test if these variables were important in item difficulty.

Consequences. Finally, the *consequences* (of test use) aspect of construct validity concerns evidence about possible adverse impact on different groups of examinees. While individual item or test scores may not show significant or impactful differential item functioning or differential test functioning, nonetheless studies may show that the test has adverse impact if used for selection or placement. Note that adverse impact could be based on construct relevant or construct-irrelevant aspects of performance. An important outcome of studies on consequences is to provide feedback to test developers. If the unintended consequences result from construct-irrelevant variables, there may be aspects of test specifications and item design that could be changed to reduce impact.

Background Aspects

As shown in Figure 5.1, preceding the five core aspects of construct validity are *practical constraints*, *latent process studies*, *conceptual framework*, and *psychometric analyses*, which, in turn, can influence *item design*, *test specifications*, and *scoring models*. Because *item design*, *test specifications*, and *scoring models* are aspects that can be most directly manipulated by test developers, it is important to consider first the background aspects that can guide these manipulations.

Practical constraints. Practical constraints can impact both the type of items that can be administered and the testing conditions. For example, equipment or testing site conditions may limit the types of items that can be administered. If testing cannot be computerized, then interactive item content, dynamic testing, or adaptive item selection are not feasible. Another important practical constraint is the duration of the testing. Longer tests and/or complex items may need to be limited to accommodate shorter testing times for example. Similarly, limited budgets for test analysis may preclude written or other **constructed response** items. Thus, practical constraints can impact the nature of the construct that can be measured through feasible item designs and test specifications.

Latent process studies. Latent process studies concern the impact of various item features, content, and testing conditions on the processes that examinees employ in responding to items. Studies that employ eye-trackers, videos, or concurrent and retrospective reports may aid in elucidating these processes. Latent process studies can concern items as well as the test as a whole. For example, if the stimulus features that are hypothesized to impact cognitive complexity differ between test items, then studies on the relationship of the features to item difficulty or response time provide evidence about

the hypothesized processes. If the relationship is strong, prediction of item psychometric properties from stimulus features may be feasible. Alternatively, consider that long and complex item types may place heavy demands on working memory or may require the development of strategies for item solving. Similarly, tests with redundant item features may result in more automated processes.

Latent process studies are also relevant to item generation. Item stimulus features often differ systematically between different generating item structures. Thus, the relationships of item structure differences to differences in item difficulty and response time may provide relevant evidence about the hypothesized processes. While it is often assumed that the content that is sampled into the structures from databases differs randomly, evidence from studies may be needed to support this claim. For example, in generating mathematics word problems, the names of the person characters are often sampled. However, evidence may be needed to determine if less familiar names result in increased problem difficulty.

Conceptual framework. The conceptual framework refers to articulations of the background theory about the domain to be measured as relevant to the goals of measurement. For achievement tests, conceptual frameworks often result from panels of experts. Current guidelines for the design of K-12 high-stakes achievement tests, which specify the content areas and skills that should be represented to measure achievement at various levels of competency, represent an evolution of the conceptualization of standards for achievement. For example, a common guideline is the *Common Core State Standards*, which target more complex competencies along with more foundational knowledge, skills, and abilities. This impacts the evidentiary requirements for understanding cognitive complexity factors that affect item design, associated responses processes, and, thus, overall construct validity.

For trait measurement, the conceptual framework may be driven by a cognitively-grounded theory about responding in the domain. Current theories of (fluid) intelligence, for example, emphasize the critical role of working memory and control processes on responding (Shipstead & Engle, 2012). Thus, to measure intelligence, the tasks should be designed to minimize the impact of prior knowledge and should be sufficiently complex as to require both working memory and control processing. For other psychological domains, a structure of content areas, rather than a specific cognitively-grounded domain theory, may guide conceptualizations. For example, in personality measurement, John and Srivastava (1999), conceptualize content facets nested within each of the Big 5 personality traits.

Psychometric analyses. As with any testing endeavor, initial psychometric analyses of items are important in guiding the test development process. The items may have been administered on previous test forms or may be new items for initial empirical tryout. Analyses to identify items with inappropriate properties for the measurement goals of interest is relevant to both test specifications and item design, particularly if coupled with an understanding of the sources of cognitive complexity of these items. Analyses of global test properties such as structural dimensionality and score reliability, provide relevant information for choosing and refining scoring models from frameworks such as CTT or IRT.

Test Development Activities

As mentioned in the previous sub-section, within the construct validity framework shown in Figure 5.1 are three aspects that can be most directly manipulated by the test developer to impact construct validity, which are *item design*, *test specifications*, and *scoring models*. Each of these aspects has direct impact on at least one aspect of validity and indirect impact on the other aspects; I discuss each of these aspects in turn in this sub-section.

Test specifications. The representation of item features on the test and the conditions of testing constitute test specifications. As shown in Figure 5.1, the *test specifications* aspect has a direct impact on the other two aspects that test developers can manipulate, *item design* and *scoring models*, as well as on the *content* and *response processes* aspects of construct validity. As discussed before, test specifications can include the proportional representation of the targeted skills or attributes on the test and their complexity or difficulty levels. For example, blueprints specify the relative representation of various item content features on the test that are assumed to involve the target skills or attributes.

Current achievement blueprints for high-stakes tests are often quite detailed and specific about the target skills and their representation, for example as found in current state achievement tests based on *Common Core State Standards*. While blueprints for psychological tests may be less precise or even consist only of desired levels of item difficulty, the relative representation of content features nonetheless impacts the construct. Also, test specifications should include instructions, guidelines for testing conditions, time limits, and so forth. Such conditions have long been known to impact performance on cognitive tests, as they also impact self-report measures (Stone et al., 2000).

As shown in Figure 5.1, the *test specifications* aspect is generally impacted by all four background aspects of construct validity. First, *practical constraints* impact the mode of testing (i.e., computerized or paper), test length, scoring automaticity, and the nature of the instructions. Second, *latent process studies* impact the representation of item features and the design of the test instructions. Third, the *conceptual framework* generally has a major impact to assure the representation of features is consistent with the measurement of the intended construct(s). Fourth, *psychometric analyses* aid the identification of item features that do not produce desired empirical item properties.

Item design. Item design principles are directly impacted by test specifications, should be consistent with the background variables, and have a direct impact on the response processes aspect of construct validity. Item design principles traditionally include item format and some guidelines about permissible features. However, explicit inclusion of item features that impact **cognitive processes** could result in greater impact on responses processes. Although traditional item design has been more an artistic process than a rule-based process, precise item designs have a long history (Hively, Patterson, & Page, 1968; Roid & Haladyna, 1982). The importance of precise specifications has been apparent more recently with the advent of item generation (Gierl & Haladyna, 2013; Irvine & Kyllonen, 2002).

Haladyna and Rodriquez (2013) describe several approaches in which **item templates** and **item shells** are used to define item structures for algorithmic item generation. Computerized approaches involve even more precise item designs as computer programs

generate items using abstractly defined structures into which content from specified databases, along with sampling rules, determine the exact item content. Cognitive psychology principles are important both in designing structures from which many items can be generated and in defining databases for permissible substitutions into the structure.

The conceptual framework and latent process studies should have major impact on item design activities to assure the measurement of the intended construct(s). However, practical constraints also must be considered in item design. For example, in a computerized test, it is possible to include dynamically designed items in which item content or feedback depends on examinee responses. However, such items are not effectively administered by paper and pencil mode. Similarly, **automated scoring**, as currently available, limits item response formats to certain kinds. For example, an automated scoring of extended constructed response items that are automatically generated is generally not feasible even though some advances have been made for certain domains such as mathematics.

Scoring models. The test developer must select a scoring model for the test, which impacts internal structure directly. Scoring may be specified as either unidimensional or multidimensional under either CTT or IRT approaches. Relevant findings from psychometric analysis may impact the choice of the eventual scoring model. Also important are test specifications in setting the limits of the heterogeneity of item content, which impacts the appropriateness of unidimensional versus multidimensional scoring models. However, practical constraints, as implemented in test specifications, can again also be important. Many test administration algorithms include only the most basic scoring algorithm focused on total scores and item difficulty so that IRT scoring models that include an item discrimination index for each item will not be feasible if immediate feedback on scores is required.

Impact and Feedback

Cognitive variables, as represented in the *latent process studies* and the *conceptual framework*, aspects, can impact all aspects of validity if they are considered and implemented through test development activities, as described above. That is, cognitive variables impact directly *item design* and *test specifications* aspects, which in turn impact the five core aspects of validity. Finally, it should be noted that the construct validity framework includes feedback loops from the four background aspects and five core aspects of validity to the three design activities that test developers can most directly manipulate. Therefore, evidence from *response processes*, *internal structure*, *relationship to other variables*, and *consequences* aspects may either support the original item and test designs or indicate needed changes. However, importantly, the external relationships do not define what is measured.

Example: Test Form Development with Generated Items

In this section I discuss the development of a new form for a test of fluid intelligence, the *Abstract Reasoning Tests* (ART), to illustrate the interrelationships of many aspects of the integrated construct validity framework, especially with respect to the types of

empirical evidence collected. This new form, the ART-E1, was intended for law enforcement personnel selection. Both fixed test forms and an item bank for adaptive testing had been developed for ART generally but the ART-E1 specifically was produced via automated item generation and hence reflects explicit item design based on cognitive principles.

Background Aspects

Conceptual framework. The main intended construct measured by the ART tests is fluid intelligence, as explicated in the Cattell-Horn-Carroll theory (see McGrew, 2005). The ART tests emphasize inductive and deductive reasoning processes by requiring the examinee to infer relationships and apply rules to non-verbal items. Figure 5.2 presents a sample item in which the examinee must select the response option that completes the rules in the 3×3 matrix.

Since the items contain no verbal content, the impact of acquired knowledge and vocabulary is minimal. Anticipated uses include personnel selection, educational selection, and placement and research studies. Similar tests, such as the *Advanced Progressive Matrices* (APM) (Raven, Raven, & Court, 2003), have been used for a variety of purposes, including cross-cultural comparisons of intelligence.

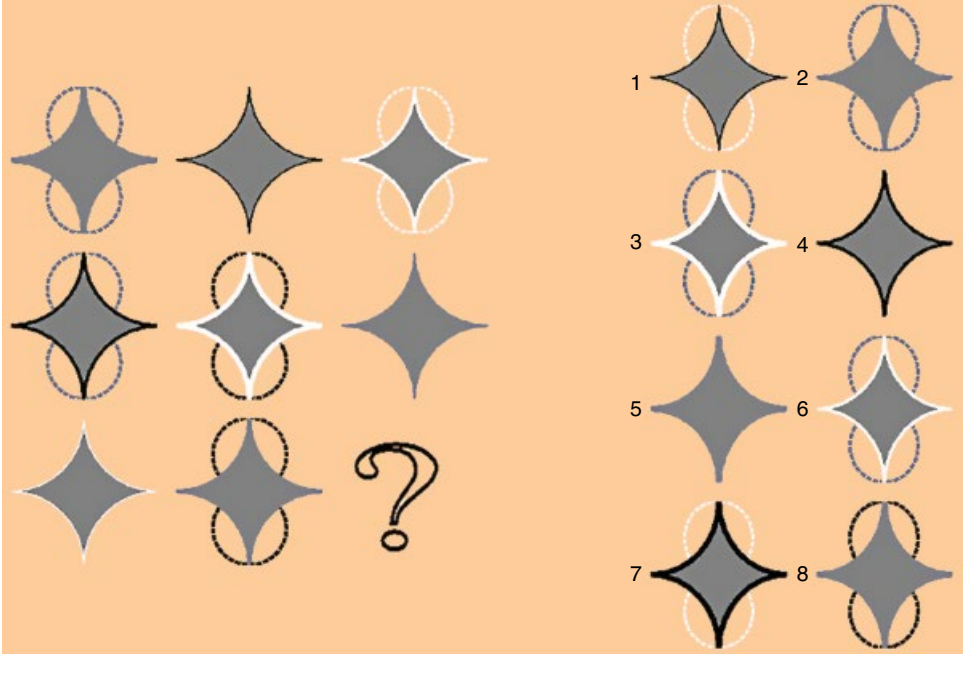

Figure 5.2 Item from ART.

The conceptual framework of fluid intelligence was deemed desirable for a test to be used for selecting law enforcement personnel, which was to be evaluated against the currently used *Cattell Culture Fair Intelligence Test*. A major aspect of test validity was predicting success in completing required training modules. Due to the diverse applicant population, minimal adverse impact for African-Americans and Hispanics was highly desirable and the general item difficulty level needed to be appropriate for the diverse applicant pool.

Latent process studies. Cognitive studies on matrix completion problems were conducted during the early development of both the ART (Embretson, 1999) and the APM (Carpenter, Just, & Shell, 1990). The Carpenter et al. (1990) theory postulates that item solving involves sequential processes, consisting of *encoding elements*, *comparing successive objects*, *inferring rule instances*, *inferring general rules*, and *applying rules to deduce the missing element*. These processes involve both executive control processes and working memory capacity.

Two variables that were salient in this research were the *number of rules* and the *level of rule abstractness* of items. The five types of rules under consideration for item design, ordered by level of abstraction, are the following:

1. *Constant in a Row* (i.e., the same property or figure appears across row or columns)
2. *Pairwise Progressions* (i.e., a property or figure changes in the same manner across rows or columns)
3. *Figure Addition/Subtraction* (i.e., adding or subtracting the first two entries in a row or column leads to the third entry)
4. *Distribution of Three* (i.e., properties or figures occur once in each row and column), and
5. *Distribution of Two* (i.e., a distribution of three with a missing entry).

Carpenter et al. (1990) postulated that lower-level rules with less complexity are attempted sequentially by examinees prior to higher-level rules when solving an item.

Carpenter et al.'s basic variables, plus some additional perceptual/display features, were implemented to predict empirical item difficulty for both APM (Embretson, 2002) and ART (Embretson, 1999, 2002). The regression coefficients for a more recent parsimonious model to predict IRT-calibrated difficulties for 150 ART items are shown in Table 5.1. The memory load variable in Table 5.1 is a combination of Carpenter et al.'s (1990) "rule levels" with the "number of rules." That is, the sum of the rule levels in an item represents memory load requirements for item processing. The analyses showed that memory load alone strongly predicted item difficulty ($R = .758$, $F = 99.472$, $p < .001$). The four perceptual variables added significantly to prediction ($F = 4.508$, $df = 3, 144$) with a final multiple correlation of $R = .782$. Similar results were obtained for response time modeling providing further support for the cognitive complexity model.

Table 5.1 Prediction of item difficulty in generated bank of 150 ART items.

	b	SE_b	β	*t*	*p-value*
(Constant)	−2.822	.302		−9.332	.000
Memory Load	.199	.019	.601	10.664	.000
Unique Elements	.172	.044	.225	3.923	.000
Object Integration	.387	.168	.129	2.301	.023
Distortion	.507	.260	.105	1.953	.053
Fusions	−.279	.185	−.084	−1.508	.134

Practical constraints. A primary practical constraint for the new form of ART was test administration by paper and pencil. Thus, a fixed test content was needed, with item difficulties appropriate for the target population of young adults applying for law enforcement positions. Further, scoring should be possible without computerization so that immediate feedback on total correct could be given to examinees upon completing the test.

Psychometric analysis. Generated items for ART had been previously administered to a population similar to the law enforcement applications. That is, 789 Air Force recruits had been administered one of three forms of ART containing 30 items. Each test form included four linking items and 26 items that were generated from the same item structure. Each structure occurred in the same test position on each form but the exact variant differed between forms. Since ART-E1 was to be generated from many of these structures, an initial analysis of these items provides important expectations for new items.

In one set of analyses, the 2PL model mentioned previously was estimated under constrained and free conditions. In the free condition, the parameters for the 3 forms × 26 items = 78 variants were freely estimated and the parameters of the four linking items that were common across test forms were constrained to equate all item parameters across groups. This condition appeared to fit the data well (−2lnL = 27,700.10; AIC 28,084.10) as the χ^2 statistics for item fit were statistically significant ($p < .01$) for only 3 items.

In the constrained condition, the items for each generating structure were constrained across test forms; hence only 26 unique items + 4 linking items = 30 item parameters across forms were estimated. This condition did not fit the data quite as well as the free condition (−2lnL = 28,389.39; AIC = 28,533.39) and the difference between models was statistically significant ($\Delta\chi^2 = 689.29$; $\Delta df = 52$; $p < .001$). However, item fit differed only slightly and the χ^2 statistics for item fit were statistically significant ($p < .01$) for 10 items. Furthermore, the trait level estimates correlated highly across conditions ($r = .969$). There was a slight impact on the distribution of the trait level estimates based on the parameters from the constrained condition (M = .108, SD = 1.00) versus the free condition (M = .02, SD = .994). The empirical reliability of trait level estimates differed little between the constrained condition ($r_{tt} = .837$) and the free condition ($r_{tt} = .853$). Similar results were obtained based on **Rasch-model** estimates obtained under free and constrained conditions.

In general, these data suggest that ART-E1 items, if generated from the same item structures, would have adequate psychometric properties. Further, the comparisons between the free and constrained estimation conditions support using item parameters based on item structures. As previously elaborated (Embretson, 1999), using the predicted

parameters rather than newly calibrated parameters results in little loss of precision in estimating person trait levels. Consequently, little or no item tryout may be needed to obtain parameter estimates for each generated item.

Test Development Activities

Item design. The conceptual framework for ART was implemented with an *automatic item generator* (Embretson, 2002, 2007). Item generators generally require the development of underlying item structures and databases to fill the structures with content. For ART, the item structure consisted of an abstract representation of the rule pattern in a matrix plus the nature of the display of objects. Several such item structures were embedded in the automatic item generator, along with databases of figures and attributes; Figure 5.2 presents such a generated item.

The item structures had a strong prediction of item difficulty ($R = .90$) in the original ART item bank meaning that item structure differences were well able to account for differences in observed item difficulties. Importantly, the cognitive complexity variables may be directly derived from the item structures thus permitting the anticipation of the level and sources of cognitive complexity in previously untried item structures. For ART-E1, new items were generated from both previously and newly implemented item structures. All previously implemented item structures had eight response options but many structures were revised to contain only six response options to reduce item difficulty.

Test specifications. The main test specifications for the ART-E1 concerned the relationships between the items and the display properties. Specifically, to maximize working memory demand, it was specified that items should have at least two relationships. Further, it was decided that items with four or more relationships should be included and that the types of relationships embedded in the item structures should vary to maximize demands on inductive reasoning. Including items that involve only pairwise progression relationships, for example, would not involve the complex processing of inferring the nature of the relationships. Also, it was decided that the perceptual properties should vary by including items with either integrated or separated object displays as well as by including items with objects that are distorted. Another item specification was derived from the practical constraints. That is, to minimize testing time and to assure that items are not too difficult, it was decided that most items on the ART-E1 should include six, rather than eight, response options. Finally, another important aspect of test specifications concerned the instructions that preceded the test. To maximize deductive and inductive processes, it was decided that the goals of problem solving should be clearly described to examinees. To accomplish this, the original ART instructions were revised to include examples of each type of relationship.

Scoring models. The practical constraints of paper and pencil testing, combined with immediate feedback to determine eligibility, dictated that test interpretations based on total scores were needed. Thus, either the CTT approach based on total test scores or IRT-based scoring with the Rasch model was possible; Rasch model scoring was

eventually implemented with conversion tables of total scores to IRT scores due to the close theoretical relationship between the two under these models.

Aspects of Construct Validity

Content. The content of ART-E1 items was defined by the variables that predict the difficulty levels and sources of cognitive complexity; Table 5.2 presents the representation of the cognitive complexity variable, memory load, and the major perceptual display variables, object integration (separated vs. integrated) and distortion (yes vs. no) in the 32 items for ART-E1. It can be seen that a broad range of items with different amounts of memory load was included. Finally, as shown in Table 5.2, both types of object integration were included at most levels of memory load, except for very high memory load levels in which several objects are needed to operationalize object relationships.

Specifically, as shown in Table 5.3, the mean number of relationships in items is about 3. For relationship types, approximately equal percentages of *Constant in a Row* (17%) and *Pairwise Progressions* (16%) were embedded in items. The higher-level relationships, *Distribution of Three* (35%) and *Distribution of Two* (26%) were more frequent, consistent with emphasizing inference of more abstract relationships. *Figure Addition/Subtraction* relationships (6%) were included only on the equated items from a previous version of the item generator. Table 5.3 also presents additional descriptive statistics on item content. The mean number of unique elements in items (to carry relationships) was approximately 3 and a large proportion of items had six response options. Slightly more than 25 percent of the items had relationships based on distorted objects while relatively few items had fusions of objects in which case it is not possible to confirm all the relationships in an item. Finally, preceding the test, instructions about the five types of relationships were presented; specifically, each type of relationship was presented with an example and then a practice item with associated diagnostic feedback.

Table 5.2 Number of items by memory load and object display type.

Memory load	*Object integration*		*Distortion*		
	Separated	*Integrated*	*No*	*Yes*	*Total*
5	2	3	3	2	5
6	0	1	1	0	1
7	1	1	2	0	2
8	1	3	0	4	4
9	1	2	3	0	3
10	3	1	4	0	4
11	4	1	4	1	5
12	0	1	1	0	1
13	2	1	3	0	3
14	1	0	0	1	1
15	1	0	0	1	1
19	2	0	2	0	2
Total	18	14	23	9	32

Table 5.3 Descriptive statistics for item features in generated ART-E1 items.

	Mean	*Standard deviation*
Number of Relationships	3.13	.942
Memory Load	9.66	3.810
Number of Unique Elements	3.09	1.329
Object Distortion	.28	.483
Object Integration	1.60	.496
Fusion	.13	.336
Number of Six Option Items	.69	.471

Response processes. ART-E1 was administered for empirical tryout to a sample of 444 law enforcement applicants. Responses processes were studied by modeling empirical item difficulty estimates from the Rasch model using two methods: (1) indirectly, predicting item difficulty using the weights for the cognitive complexity variables in Table 5.1 based on studies of previously generated items, and (2) directly, predicting item difficulty with weights estimated for the cognitive complexity variables in Table 5.1 from the current data.

To apply the indirect method, the calibrated Rasch item difficulties on ART-E1 were equated to the original ART item bank through eight common items. The estimated item difficulties for these items correlated strongly between the two sources ($r = .856$); thus, an equating constant was estimated ($c = -.0977$) and applied to all ART-E1 items. The equated item difficulties were then compared to the predictions from the cognitive model developed for previously generated ART items, as described earlier. Both prediction methods required an additional adjustment for the ART-E1 items generated with six options, rather than with eight options. This was done by adding a binary variable (*ALT6*) to the indirect prediction model that classified the items by whether they had six or eight response options. The regression of ART-E1 item difficulties on predicted item difficulties was strong ($R = .719$, $p < .001$), with the prediction given as follows: $\beta'_{ART\text{-}E1} = .705(\beta'_{indirect}) - 1.473(\text{ALT6})$. Thus, the cognitive model based on the original bank of generated items is supported for the generated items on ART-E1.

For the direct prediction method, the item difficulty parameters from the Rasch model calibrated directly from ART-E1 data were regressed on the item complexity scores obtained from the item generation specifications. Estimating new weights for the cognitive model resulted in negligible changes in prediction. The five cognitive complexity variables (i.e., *Memory Load, Number of Unique Elements, Fusion, Distortion,* and *Object Integration*) had a strong relationship to item difficulty ($R = .706$, $p < .001$), with *Memory Load* as the strongest predictor. Adding *ALT6* to reflect item format changes led to somewhat stronger prediction ($R = .773$, $p < .001$) as compared to the indirect prediction model. As for the indirect prediction method, *Memory Load* and *ALT6* were the strongest predictors. Thus, the dominant impact of memory load in item difficulty was supported for ART-E1. In sum, in the context of the background research on latent processes, ART-E1 is supported as measuring fluid intelligence with the same sources of cognitive complexity as the original ART demonstrating the powerful capability of the automated item generator to produce items with predictable levels and sources of cognitive complexity.

Internal structure. Scoring models were compared to determine the possible loss of information by not including item discrimination weights using the previously used data on ART-E1 for 444 law enforcement applicants. Two IRT models were specifically fit to the data: the Rasch model discussed in the previous sub-section (−2lnL = 13,212.62; AIC = 13,278.62) and the 2PL model (−2lnL = 13,073.49; AIC = 13,201.49). Since these are nested models, a Chi-square difference test was conducted and was found statistically significant ($\chi^2 = 139.13$; $df = 32$, $p < .001$) showing that the 2PL model represents an empirical improvement in model-data fit over the Rasch model. An inspection of marginal χ^2 item-level fit statistics indicated one misfitting item ($p < .01$) for the 2PL model and two misfitting items for the Rasch model ($p < .01$); thus, item fit was not substantially different between the two models. Further, the regression of 2PL trait levels on Rasch trait levels indicated nearly identical estimates with unstandardized prediction weights close to a perfect transformation ($R = .986$; $b = .993$; $a = .000$).

Thus, while the 2PL model did fit the data better overall, item fit and trait level estimates differed little between models and the Rasch model was supported for scoring; for the Rasch trait level estimates (M = .0002, SD = .890) the empirical reliability of scores was .809. Differential item functioning was examined next, comparing African-American and Hispanic applicants to a reference group of Caucasian applicants. Using the full item set as anchors, the χ^2 statistics for the item parameter contrasts across groups indicated no statistically significant effects ($p > .01$).

CTT statistics were also estimated. The Cronbach's alpha internal consistency estimate was .792 and raw test scores were relatively high (M = 21.460, SD = 4.782). Classical item statistics indicated a mean item *p*-value of .671 and a mean biserial correlation of .557 similarly reflecting items of above-average difficulty and moderate discriminatory power. A further analysis was conducted to determine the impact of the cognitive complexity factors on item discrimination. With the ART-E1 biserial correlations as a dependent variable, the only statistically significant predictor was *Fusion* ($r = -.402$, $p = .010$); it should be noted that three of the four items with *Fusion* had marginal fit in the Rasch model ($.01 < p < .05$). Taken together, these results also support the internal structure of the test and also provide feedback for changes in test specifications. For example, while item discrimination appears to be generally acceptable, eliminating items with fusion would likely increase discrimination and hence internal consistency. Further, adding some more difficult items would also impact internal consistency if more *p*-values were close to .50.

Relationship to other variables. In a separate study, the utility of ART-E1 for the prediction of classroom training success for 152 law enforcement recruits was investigated. Specifically, 12 training modules concerning aspects of law enforcement duties were completed by all recruits. Following each module, a series of **multiple choice** items were administered to assess learning and the mean score across modules was computed to assess overall training outcomes.

All recruits had been previously administered three tests as potential predictors of training success, which were the ART-E1, the *Cattell Culture Fair Intelligence Test III* (CFIQ) and the *Reasoning* factor (*B* factor) from the *16 Personality Factor* test (Cattell, Cattell, & Cattell, 1993). All three tests correlated significantly with training success: ART-E1 ($r = .333$, $p < .000$), CFIQ ($r = .211$, $p = .009$), and *B* factor ($r = .231$, $p = .004$);

however, the ART-E1 correlation with training success was significantly higher than those of the other two tests ($p < .001$).

Consequences. ART-E1 was being considered as a replacement for the CFIQ to select applicants for law enforcement positions. Under current use of the CFIQ, at the first stage of personnel selection, applicants are screened out for low scores (i.e., bottom 10%) on the CFIQ. An important consideration in such a situation is the potential impact of the test use on racial-ethnic diversity in the law enforcement officers that are eventually hired.

Figure 5.3 presents the cumulative percentage of raw scores on both ART-E1 and the CFIQ for the sample of 444 applicants that I discussed earlier. For the CFIQ, the cut point of 10% leads to larger percentages of African-American and Hispanic applicants eliminated from further consideration as compared to Caucasian applicants; this was similarly true for all higher cut points. In contrast, for the ART-E1 there are only negligible differences in the sample composition in terms of racial-ethnic groups at the cut point, supporting no adverse consequences of test use in terms of ethnicity. At higher cut points, some differences between ethnic groups were observed, but the gaps were not as large as for the CFIQ. Thus, the data on consequences support the conceptual framework and goals of measurement.

One meaningful question to ask is why the ART was associated with lesser adverse impact. The most salient hypothesis is that the extended instructions of the ART reduced racial-ethnic differences in test-wiseness. The instructions for the CFIQ are very brief, with little explanation about the possible relationships that would be regarded as appropriate. With ART, the instructions included an example of each type of relationship that could occur. These instructions established the rules of item solving, but, of course, the examinees must still infer the relationships in each item and reach the appropriate conclusion. The best test of this hypotheses about the differential impact of instructions would be an experiment with ART, with random assignment to the current instructions versus brief instructions. Unfortunately, experimentation is difficult in the context of an operational testing program.

Summary: Construct Validity for the Automatically Generated Test

The items for ART-E1 were produced from an automatic item generator. Importantly, the item structures in the generator could be linked to cognitive complexity variables that had been studied on previously generated items. Thus, the *item design* and *test specifications* aspects could be based explicitly on cognitive complexity variables. Also included in the test development process was the *test specifications* aspect based on prior research with scoring models.

The studies presented in this chapter lent empirical support to all five aspects of construct validity for ART-E1. Specifically, the evidence for the *content* aspect of construct validity supported the composition of test content and involved diverse and complex relationships presented under varying perceptual displays. The evidence for the *response processes* aspect of construct validity was driven by item difficulty modeling, with strong relationships produced by weights estimated from either the data or previous studies; memory load was the predominant factor in cognitive complexity. The evidence for the

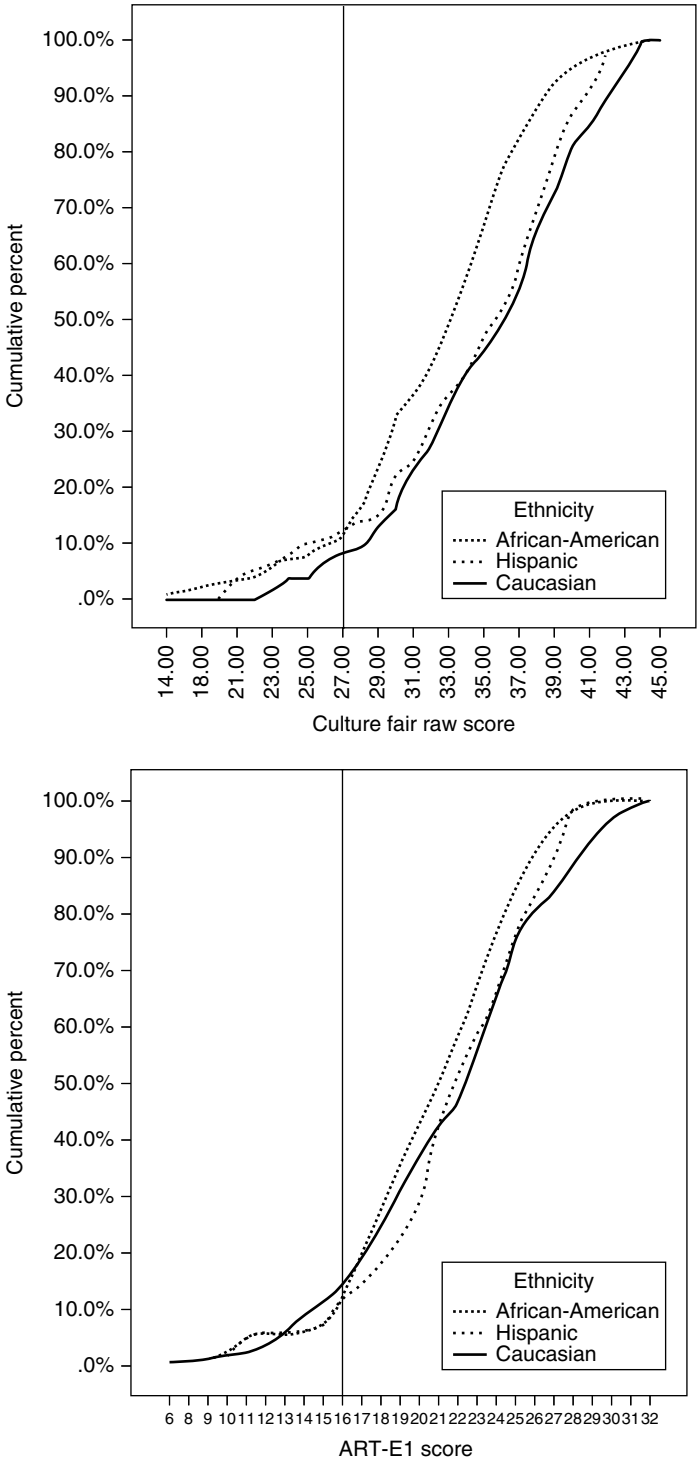

Figure 5.3 Cumulative percentages of scores on the ART-E1 and CFIQ by race-ethnicity.

internal structure aspect of construct validity, which includes empirical estimates of reliability, was gathered using the desired scoring models. The evidence for the *relationship to other variables* aspect came from correlations of ART-E1 test scores with scores from a course completion measure that were stronger that then correlations of test scores from other tests, especially the one currently in use. Finally, the reduced adverse impact on the ethnic composition of test takers that are selected under different cut scores provided evidence for the *consequences* aspect of construct validity.

Interestingly, the potential of an operational use of an automatic item generator was also strongly supported through this work, especially when coupled with a strong research foundation on cognitive complexity factors that affect response processing. That is, ART-E1 consisted of both new and previously used item structures, both of which could be linked to cognitive complexity variables that had been established as predictors of item difficulty. The results for predicting ART-E1 item difficulties from previous developed weights was sufficiently strong as to support minimal requirements for the tryout of new items.

Conclusion

The purpose of this chapter was to provide a general framework for construct validity that could be broadly applicable to the incorporation of cognitive principles in testing. The framework was explicitly based on the five core aspects of construct validity in the current *Standards* and focused on the interrelationships and integration of evidence for different aspects of construct validity. Background aspects, including evidence from foundational research, were shown to have important potential impact on the design activities that test developers can most directly manipulate.

An example of the development of a new form for a test of fluid intelligence, the ART-E1, was presented to illustrate the relationships between the evidence collected for background aspects, test development aspects or activities, and the five core aspects of construct validity. Specifically, an automatic item generator had been developed for the test with item structures that could be explicitly linked to previously established sources of cognitive complexity.

The data presented supported both the new test and capacity of the item generator to produce items with predictable properties in the validity system. If results on different types of tests with other item generators were to be similarly supported, the burden of test development activities could be substantially reduced for such tests in the future in that automatically generated items probably could be used with little or no tryout.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bejar, I. I. (2002). Generative testing: From conception to implementation. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 199–218). Mahwah, NJ: Erlbaum.

Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals*. New York, NY: McKay.

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer-Verlag.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of processing in the Raven's Progressive Matrices Test. *Psychological Review*, *97*, 404–431.

Cattell, R. B., Cattell, A. K., & Cattell, H. E. P. (1993). *16PF Fifth Edition Questionnaire*. Champaign, IL: Institute for Personality and Ability Testing.

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

Daniel, R. C., & Embretson, S. E. (2010). Designing cognitive complexity in mathematical problem solving items. *Applied Psychological Measurement*, *34*(5), 348–364.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: The Guilford Press.

Defense Manpower Data Center (2008). *Manual for the assembling objects tests*. Monterey, CA.

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

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika*, *64*, 407–433.

Embretson, S. E. (2002). Generating abstract reasoning items with cognitive theory. In S. Irvine & P. Kyllonen (Eds.), *Generating items for cognitive tests: Theory and Practice* (pp. 219–250). Mahwah, NJ: Erlbaum.

Embretson, S. E. (2007). Construct validity: A universal validity system or just another test evaluation procedure? *Educational Researcher*, *36*(8), 449–455.

Gierl, M., & Haladyna T. M. (Eds.) (2013). *Automatic item generation: Theory and practice*. New York, NY: Routledge.

Gierl, M., Zhou, J., & Alves, C. (2008). Developing a taxonomy of item model types to promote assessment engineering. *Journal of Technology, Learning and Assessment*, *7*(2), 4–50.

Gorin, J. (2006). Item design with cognition in mind. *Educational Measurement: Issues & Practices*, *25*(4), 21–35.

Goto, T., Kojiri, T., Watanabe, T., Iwata, T., & Yamada, T. (2010). Automatic generation system of multiple-choice cloze questions and its evaluation. *International Journal on Knowledge Assessment and E-Learning*, *2*(2), 210–224.

Haladyna, T. M. & Rodriquez, M. C. (2013). *Developing and validating test items*. New York, NY: Routledge.

Hively, W., Patterson, H. L., & Page, S. (1968). A "universe-defined" system of arithmetic achievement tests. *Journal of Educational Measurement*, *5*, 275–290.

Hoyt, C. (1941). Test reliability estimated by analysis of variance. *Psychometrika*, *6*, 153–160.

Irvine, S., & Kyllonen, P. (Eds.) (2002). *Item Generation for Test Development*. Mahwah, NJ: Erlbaum.

John, O. P., & Srivastava, S. (1999). The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (Vol. *2*, pp. 102–138). New York, NY: Guilford.

Leighton, J., & Gierl, M. (2007a). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge, UK: Cambridge University Press.

Leighton, J., & Gierl, M. (2007b). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues & Practices*, *26*(2), 3–16.

Lord, F. M., & Novick, M. R. (with contributions by Allan Birnbaum) (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Luecht., R. M. (2013). An introduction to assessment engineering for automatic item generation. In M. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 59–75). New York, NY: Routledge.

McGrew, K. S. (2005). The Cattell-Horn-Carroll theory of cognitive abilities: Past, present, and future. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 136–182). New York, NY: Guilford.

Mortimer, T., Stroulia, E., & Yazdchi, M. (2013). A web-based automatic item generator. In M. Gierl & T. Haladyna (Eds.), *Automatic item generation* (pp. 217–230). New York, NY: Routledge.

Newstead, S. E., Brandon, P., Handley, S. J., Dennis, I., & Evans, J. S. (2006). Predicting the difficulty of complex logical reasoning problems. *Thinking & Reasoning*, *12*(1), 62–90.

Raven, J. C., Raven, J., & Court, J. H. (2003). *Manual for Raven's Progressive Matrices*. San Antonio, TX: Harcourt Assessment.

Rijmen, F., & DeBoeck, P. (2001). Propositional reasoning: The differential combination of "rule" to the difficulty of complex reasoning tasks. *Memory & Cognition*, *29*(1), 165–175.

Roid, G. H., & Haladyna, T. M. (1982). *A technology for test-item writing*. New York, NY: Academic Press.

Rupp, A., Templin, J., & Henson, R. (2010). *Diagnostic assessment: Theory, models and applications*. New York, NY: Guilford.

Singley, M. K., & Bennett, R. E. (2002). Item generation and beyond: Applications of schema theory to mathematics assessment. In S. Irvine & P. Kyllonen (Eds.), *Item generation for test development* (pp. 361–384). Mahwah, NJ: Erlbaum.

Shipstead, Z., & Engle, R. W. (2012). Interference within the focus of attention: Working memory tasks reflect more than temporary maintenance. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *39*(1), 277–289.

Stone, A. A., Turkkan, J. S., Bachrach, C. A., Jobe, J. B., Kurtzman, H. S., & Cain, V. S. (2000). *The science of self-report: Implications for research and practice*. Mahwah, NJ: Erlbaum.

Webb, M. L. (1997). *Criteria for alignment of expectations and assessments on mathematics and science education*. Research Monograph Number 6. Washington, DC: CCSSO.

Webb, M. L. (2002). *Technical issues in large scale assessments*. Research Monograph. Washington, DC: CCSSO.

6

The Role of Cognitive Models in Automatic Item Generation

Mark J. Gierl and Hollis Lai

Introduction

Automatic item generation (AIG) (Embretson & Yang, 2007; Gierl & Haladyna, 2013; Irvine & Kyllonen, 2002) is the process of using models to generate test **items** with the aid of computer technology. The purpose of this chapter is to describe and illustrate the important role that **cognitive models** play in the item generation process. A cognitive model for AIG is a representation that highlights the **knowledge, skills, and abilities**, and/or content required to generate new test items. To begin, we establish a context for the application of the item generation method presented in this chapter by describing why researchers and practitioners have developed a voracious appetite for test items. Next, AIG is offered as a possible solution to whet this appetite. We describe a three-step method for implementing AIG. In step 1, test developers identify the content required for item generation. In step 2, an **item model** is developed to specify where this content is placed in each generated item. An item model is similar to a template that highlights the variables in a test item that can be manipulated to produce new items. In step 3, computer-based algorithms are used to place the content into the item model. Then, we focus on how the content specified in step 1 can guide the generative process by describing two different types of cognitive models – the *logical structures* and the *key features* cognitive model. To ensure our description is practical and concrete, we draw on examples from K-12 science and medicine. Finally, we evaluate the model-based approach presented in this chapter. We also describe how outcomes from future research on cognitive modeling could enhance the item generation process.

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