13

Measuring Human Intelligence with Artificial Intelligence

Adaptive Item Generation

Susan E. Embretson

INTRODUCTION

Adaptive item generation may be the next innovation in intelligence testing. In adaptive item generation, the optimally informative item is *developed anew* for the examinee during the test. Reminiscent of computer versus person chess games, the computer generates the next item based on the previous pattern of the examinee's responses. Adaptive item generation requires the merger of two lines of research, psychometric methods for adaptive testing and a cognitive analysis of items.

Adaptive testing is the current state of the art in intelligence measurement. In adaptive testing, items are selected individually for optimal information about an examinee's ability during testing. The items are selected interactively by a computer algorithm using calibrated psychometric properties. Generally, harder items are selected if the examinee solves items, while easier ones are selected if the examinee does not solve items. Adaptive item selection leads to shorter and more reliable tests. In a sense, optimal item selection for an examinee is measurement by artificial intelligence.

Adaptive item generation is a step beyond adaptive testing. Like adaptive testing, it estimates the psychometric properties of the optimally informative items for the person. Beyond this, however, the impact of specific stimulus content on an item's psychometric properties must be known. That is, knowledge is required of how stimulus features in specific items impact the ability construct.

This chapter describes a system for measuring ability in which new items are created while the person takes the test. Ability is measured online by a system of artificial intelligence. The items that are created are designed to be optimally informative about the person's ability. The system behind the item generation is the cognitive design system approach (Embretson, 1998).

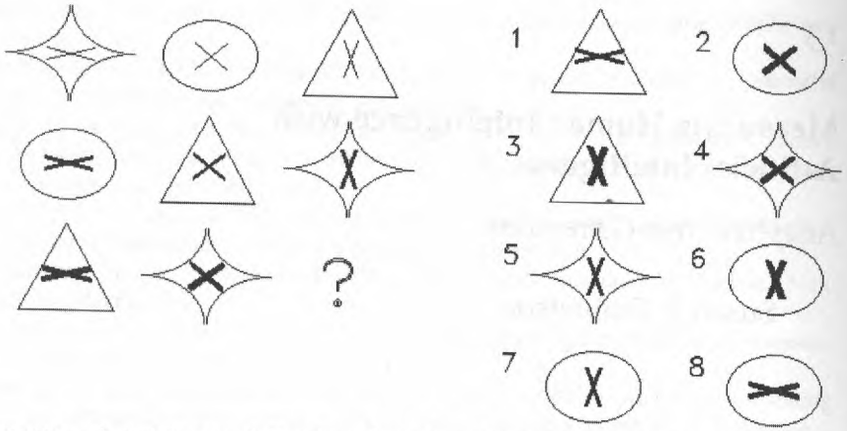

FIGURE 1. A matrix completion item from the Abstract Reasoning Test.

COGNITIVE DESIGN SYSTEM APPROACH TO ADAPTIVE ITEM GENERATION

The cognitive design system approach has been applied, at least partially, to several item types that measure intelligence and aptitude. One of the most extensive applications has been to matrix completion problems, such as shown in Figure 1. Matrix completion problems are found on many intelligence tests, including the Raven Advanced Progressive Matrix Test (Raven, Court, & Raven, 1992), the Naglieri Test of Non-Verbal Intelligence Test, and the Wechsler Intelligence Test for Children. Many scholars regard this item type as central to measuring intelligence (Carroll, 1993; Gustafsson, 1988).

Central to the cognitive design system approach is a cognitive processing model for the item type that measures the construct. However, adaptive item generation also requires several other supporting developments, which include a conceptualization of construct validity that centralizes the role of item design, psychometric models that incorporate design variables, and finally, a computer program that generates items. This section describes the theoretical rationale for cognitive design systems. Then, supporting developments will be elaborated. Finally, the stages involved in applying the cognitive design system to actually generate items will be reviewed.

Theoretical Foundations for Cognitive Design Systems

A cognitive design system is based on an information processing theory of the item type. Such theories originated with cognitive component

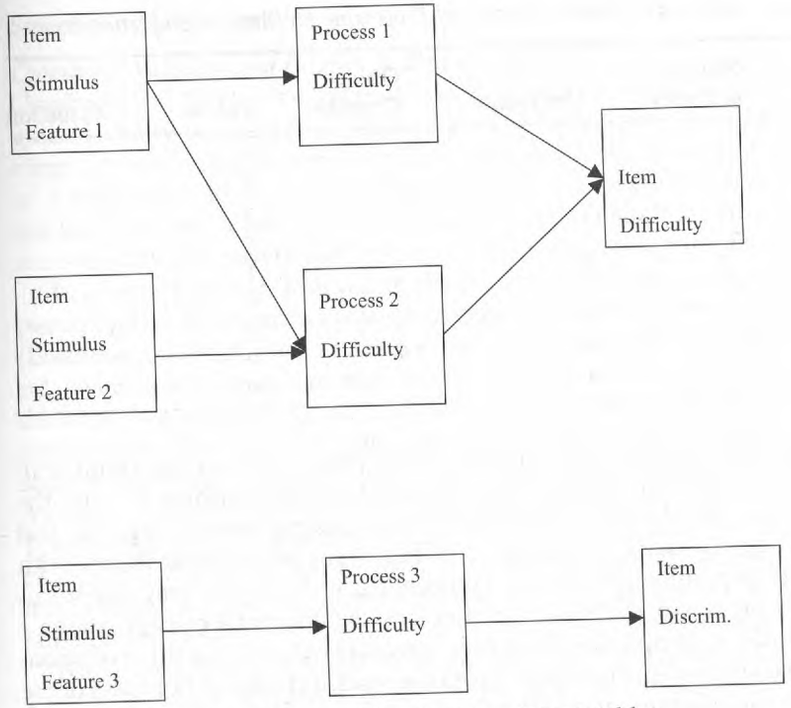

FIGURE 2. Schematic of the cognitive design system variables.

analysis of complex item types for measuring intelligence (Sternberg, 1977) or for other ability items, such as reading (Carroll, 1976). A cognitive theory specifies processes in item solution, the impact of processes on performance, and the impact of stimulus features on processes. To be useful for item generation, a primary dependent measure for performance must be item difficulty (in addition to response time) and, of course, it must be empirically supported.

Figure 2 presents the primary relationship of the cognitive theory to the psychometric properties of items. In Figure 2, the item stimulus properties are related to processing difficulty, which in turn are related to item difficulty and other item properties. Although the primary psychometric property is item difficulty, item discrimination may be influenced by the processes as well. For example, peripheral processes that are not central to the construct that is measured may lower item discrimination.

To illustrate the role of cognitive theory in psychometric tasks, consider again the matrix completion task, as shown on Figure 1. Carpenter, Just, and Shell (1990) postulated and supported two major inference processes involved in performance: goal management and correspondence finding.

TABLE 1. *Stimulus Features to Represent Processing on Matrix Completion Items*

Item	Number of Rules	Abstraction	Object Overlay	Fusion	Distortion
241	1	0	1	0	0
101	1	0	0	1	0
192	1	0	1	1	0
344	2	1	1	0	0
285	2	0	1	1	0
391	2	1	1	0	1
254	3	0	1	0	0
406	4	0	1	0	0
423	4	1	0	0	0

The stimulus features that impacted these processes were the number of rules in the problem and the abstractness of the relationships. In turn, the processes depend on the problem solver's working memory capacity and abstractness capacity. Carpenter et al. (1990) supported their theory with a variety of methods to explicate processing, including a computer simulation of processes, eyetracker studies, and experimental manipulations.

In Figure 1, completing the problem requires identifying three relationships: a change of girth of the X across the rows, a change of boldness of the X down the columns, and a distribution-of-three relationship of the outer shapes, such that each instance appears just once in each row and column. According to the Carpenter et al. (1990) theory, substantial working memory capacity is involved because lower level rules are tried before higher level rules, such as the distribution-of-three relationship. Abstraction capacity is minimized for the item in Figure 1, however, since the figures correspond directly and no entries with null values are given.

Carpenter et al.'s (1990) two major variables, number of rules and abstraction level, represent only inference processing in matrix completion problems. Encoding variables were not part of their model. However, since encoding should be included in any complex processing model, Embretson (1995b; 1998) added some variables to represent the difficulty of encoding the figures in the item. Three perceptual properties – object fusion, object distortion, and object overlay – were added to the inference processing variables for a more complete model.

Table 1 presents scores for these processing variables on some matrix completion items. If these features can be objectively scored, and in turn if they can predict both item response time and item psychometric properties, then a viable cognitive model has been developed for item generation. Although scoring matrices initially required raters, currently all features for generated items can be scored objectively from the item structure specifications (see next section).

Supporting Developments

Construct Validity and Cognitive Design Systems

Cronbach and Meehl's (1955) traditional concept of construct validity has guided ability testing for decades. It provided the conceptual underpinnings for combining diverse data about the quality of the test as a measure of a construct. However, the relevant data could accumulate only after the test was developed. Thus, the data served to elaborate the construct measured by the current test but not to provide guidance for test design.

To incorporate test design into the construct validity concept, two aspects must be distinguished: construct representation and nomothetic span (Embretson, 1983). Construct representation directly concerns the theoretical meaning of test performance. That is, construct representation concerns the processes, strategies, and knowledge that are directly involved in test performance. The research paradigm for construct representation differs sharply from nomothetic span; it involves applying cognitive psychology methods to build information processing models of the measuring task. Typical research involves manipulating the stimulus features of the task to change the relative impact of the postulated processes. This approach has implications for test design because these same features can be manipulated on test items to measure the targeted aspects of processing.

Nomothetic span overlaps substantially with the traditional construct validity concept because it concerns the empirical relationships of test scores. It provides information about the usefulness of the test for measuring individual differences. However, it differs somewhat from the traditional nomological network. That is, relationships should be predictable from construct representation.

Taken together, tests can be designed for both desired construct representation and nomothetic span. In the matrix completion problems, for example, a measure could be designed to require both working memory capacity and abstraction capacity by including matrix items that vary on both number of relationships and abstractness. Or, in contrast, the measure could be designed for only working memory capacity by excluding abstract relationships. Nomothetic span, in turn, will be influenced by these different designs. More limited empirical correlates would be expected if only one capacity was represented, for example.

Psychometric Models for Cognitive Design Systems

Cognitive design systems require psychometric models that can incorporate test design variables. This allows, item properties to be predicted from the cognitive design system variables. The state of the art in psychometric methods is item response theory (IRT). Adaptive testing typically requires IRT models to optimize item selection in measuring ability and to equate measurements between persons who are administered different sets of

items. In IRT, the probability of each person's response to each item, $P(\theta)$, is modeled from the person's ability, θ_s, and the properties of items. In the two-parameter logistic model (2PL),

$$P(\theta) = \frac{\exp(\alpha_i(\theta_s - \beta_i))}{1 + \exp(\alpha_i(\theta_s - \beta_i))}. \quad (1)$$

The item properties are item difficulty, β_i, and item discrimination, α_i. A person's ability, θ_s, is estimated in the context of a model, such as in Equation 1. The ability estimate depends not only on the accuracy of the subject's responses, but also on the parameters for the items that were administered. The item parameters are inserted into Equation 1 and ability is estimated to yield the highest likelihood of the observed responses (see Embretson & Reise, 2000, or Hambleton & Swaminathan, 1985, for more information on ability estimation).

The 2PL model, like most standard IRT models, does not include any parameters for the design features behind items. Item difficulties and discriminations are calibrated separately for each item, without regard to their specific design features. However, special IRT models have been developed to estimate the impact of design features on item difficulty (Fischer, 1973; DiBello, Stout, & Roussos, 1995; Adams, Wilson, & Wang, 1997). These models are appropriate if item discriminations do not differ or if they do not relate to design features.

The 2PL-constrained model (Embretson, 1999) was developed to allow design features to influence both item difficulty and item discrimination. Table 1 shows some scored features of matrix items that represent the cognitive model as just described. Scores for each item, q_{ik}, on the k features, define variables that can model item difficulty and discrimination. The 2PL-constrained model replaces calibrated item difficulty and discrimination with a weighted combination of the scored features, q_{ik}, as follows:

$$P(\theta) = \frac{\exp(\Sigma q_{ik}\phi_k(\theta_s - \Sigma q_{ik}\tau_k + \tau_o))}{1 + \exp(\Sigma q_{ik}\phi_k(\theta_s - \Sigma q_{ik}\tau_k + \tau_o))} \quad (2)$$

where τ_k is the parameter for the weight of feature k in item difficulty, ϕ_k is the parameter for the weight of feature k in item slope, and θ_s is the ability of person s. Notice that item difficulty and item discrimination are represented by a weighted combination of the stimulus features, $\Sigma q_{ik}\tau_k$ and $\Sigma q_{ik}\phi_k$, respectively. The weights are estimated in the IRT model to maximize fit to the item response data. Heuristically, however, the weights are roughly equivalent to regression weights in predicting item difficulty and item discrimination.

Once the weights are calibrated to reflect the impact of the stimulus features, item difficulties and discriminations for new items can be predicted directly, without empirical tryout. Obviously, reasonably accurate prediction depends on the fit of the model. An empirical example will be presented below for the matrix problems to show both the calibration of weights and the assessment of model quality.

Computer Programs for Adaptive Item Generation

Adaptive item generation requires two types of computer programs: (1) an item generator program that actually creates the items and (2) an adaptive testing program that can be interfaced with the generator. The item generator program produces items to target levels and sources of cognitive complexity in the cognitive design system approach. Item structures, which are blueprints for the items, are essential for item production. The structures carry the specific sources of cognitive complexity and predicted item difficulty. The nature of the item structures depends on the item type. For nonverbal items, the item structure determines the arrangement and display of objects. Specific objects are randomly selected to fulfill the structure. For verbal items, structures need to specify deep level meanings or logical representations that can be instantiated with different surface features, such as exact vocabulary level and syntax.

Once the item generator is developed, it then must be interfaced with an adaptive testing program. An adaptive testing program not only displays items and records responses, but also interacts with the examinee to estimate ability and to determine the optimal item properties for the next item to be administered. Several adaptive testing programs are available; however, these programs search for existing items in an item bank. To provide adaptive item generation, the testing system must be linked to item structures that produce items of target psychometric properties.

Stages in Applying Cognitive Design Systems

The cognitive design system approach may be applied to new or existing measures of a construct. The stages presented below are most appropriate for existing measures with adequate nomological span. In this case, the usefulness of the test for measuring individual differences is already established. The cognitive design system approach then can be applied to establish the construct representation aspect of construct validity. This provides a basis for designing new items and item generation, as well as possible test redesign. A new measure of construct also could be developed under the cognitive design system approach; in this case, additional studies to establish nomothetic span are needed. Also, for new measures, algorithmic item generation can occur earlier in the process.

Develop Cognitive Model for Existing Items

In the initial stages, the goal is to develop a plausible cognitive processing model for the existing ability test items. Cognitive modeling typically

begins with a literature review on underlying processing components and the stimuli that determine their difficulty. Often the literature concerns a related task, rather than the exact item type on an ability test. That is, tasks that are studied in the laboratory are often quite easy and presented in a verification format. Ability test items are much harder and presented in multiple choice format. Thus, a more complex model may need to be postulated to adequately represent processing for solving ability test items. For the matrix completion task, although Carpenter et al. (1990) studied ability test items, their model did not include encoding or decision processing. Thus, a more complete model was developed.

The next step is to empirically support the model. Data on two primary dependent variables on the ability test items must be obtained: item response time and item difficulty. These dependent variables are mathematically modeled from item stimulus features that are postulated to impact processing. Item stimulus features on existing tests often show multicollinearity, which can bias relative importance in the model. Thus, additional studies to unconfound the impact of correlated features are needed. Also, converging operations for supporting the model are needed. For example, eyetracker and simulation studies also provide information about the plausibility of the processing model.

Algorithmic Item Generation and Revised Cognitive Model

The next stage directly concerns test design; that is, can the stimulus features be manipulated separately to impact processing difficulty and item performance? To manipulate item features, a set of item specifications based on the model variables is constructed. Correlated features can be unconfounded by crossing the various levels of the stimulus features. For example, in existing matrix completion problems, the display of objects (e.g., overlay) is correlated with the number of rules. However, in algorithmic item generation, display type can be fully crossed with the number of rules and then items can be constructed to fulfill the various combinations of features.

The newly constructed items are then studied empirically to determine the impact of the stimulus design features on item performance. Although new items can be calibrated in a tryout, the main focus is on calibrating design features. The design features should be sufficiently predictive of item difficulty and other psychometric indicators, as well as response time. Items that represent the same combination of design features should be highly similar empirically.

Item Generation by Artificial Intelligence

As noted before, a computer program must be developed for item generation. Although the programming effort required to develop a mechanism to create and display items is substantial, the development of the item

structures for the particular item type is crucial to success. All items from the same structure carry the same sources and levels of cognitive complexity. Item structures can differ qualitatively between item types; therefore, a new research effort is required to develop structures that are linked to the cognitive model variables. An item generator program, ITEMGEN1, has been developed for six item types that measure nonverbal intelligence, including two types of matrix completion tasks, geometric analogies, geometric series problems, and two types of items for spatial ability (Psychological Data Corp., 2002). The structures for spatial ability items differ qualitatively from the other nonverbal intelligence item structures.

Empirical Tryout of Item Generation

The final stage involves an actual tryout of the generated items. Online testing is essential because continuous data are needed to evaluate the quality of the design principles. This stage has not yet been implemented for the cognitive design system approach.

New psychometric issues arise with adaptive item generation. First, new diagnostic indices are needed to evaluate the effectiveness of the design features, and their various combinations, in yielding items of predicted psychometric quality. Since design features probably will be contained in IRT models, as noted earlier, perhaps current indices for item fit can be extended to assess the design features. Second, further research is needed on how uncertainty in the item parameters (i.e., because they are predicted from design features) impacts ability estimates. Several studies (Bejar et al., 2002; Mislevy, Sheehan, & Wingersky, 1993; Embretson, 1999) have found that measurement error increases modestly when item parameters are predicted rather than calibrated. These studies further suggest that the impact of item uncertainty can be readily countered by administering a few more items. However, in the context of online testing, it may be possible to monitor individual examinees for the impact of item uncertainty. Research on indices to diagnose online problems is also needed.

SUPPORTING DATA FOR COGNITIVE DESIGN SYSTEMS

The cognitive design system approach has been applied to several nonverbal aptitude test items, including matrix completion problems (Embretson, 1998), geometric analogies (Whitely & Schneider, 1981), spatial folding (Embretson, 1994), and spatial object assembly (Embretson, 2000; Embretson & Gorin, 2001). A computer program for item generation has been developed for these item types.

The cognitive design system has also been applied to several other item types, including verbal analogies (Embretson & Schneider, 1989), verbal classifications (Embretson, Schneider, & Roth, 1985), letter series (Butterfield et al., 1985), paragraph comprehension (Embretson & Wetzel, 1987;

Gorin, 2002), and mathematical problem solving (Embretson, 1995a). Although a computer program for generating these items does not yet exist, research is in progress for mathematical problem solving (Embretson, 2002a) and paragraph comprehension (Embretson & Gorin, 2002; Gorin, 2002). However, since all of these item types involve words, some psycholinguistic capabilities will be required for full item generation.

In this section, empirical support for generating matrix completion problems by cognitive design systems will be described. Although research on the object assembly task for measuring spatial ability is somewhat more complete because it includes empirical tryout of AI-generated items (Embretson, 2000), matrix completion problems are often regarded as central to measuring intelligence.

Initial Cognitive Model for Matrix Items

The initial modeling for the matrix items, although conducted early in the studies on matrices, was reported in Embretson (2002b). The Advanced Progressive Matrices (APM, Raven et al., 1992) was selected as the target for an initial cognitive model for the matrix completion task for two reasons. First, Carpenter et al. (1990) had studied APM intensively to develop their theory and they provided scores for many items. Second, APM is widely recognized as a measure of fluid intelligence.

Two mathematical models for the APM were developed to begin the cognitive design system process (see Embretson, 2002b). Model 1 contained Carpenter et al.'s (1990) processing variables for rule induction, the number of rules, and the abstractness of the rules. Three variables to represent encoding difficulty were also included in Model 1. Model 2 contained an alternative measure of rule induction processing, memory load, as well as the encoding variables. The memory load variable operationalized Carpenter et al.'s (1990) postulated processing sequence of rules in matrices. That is, they postulated that examinees attempted to relate matrix entries by higher level rules only after lower level rules failed. Thus, greater amounts of processing and working memory are required for items with higher level rules because lower level rules had to be tried and remembered. Embretson's (2002b) memory load variable operationalized rule induction processing by summing the levels of the rules in each item. Both Model 1 ($R^2 = .79$, $p < .01$) and Model 2 ($R^2 = .81$, $p < .01$) provided adequate prediction of APM item difficulty. Although the encoding variables did increase prediction, the rule induction variables had the strongest impact on item difficulty.

Algorithmic Item Generation and Revised Cognitive Model

The models just identified provided the basis for constructing new items. A bank of 150 items contained five replicates of thirty item structures that

TABLE 2. *Comparisons of Alternative Psychometric Models for 90 Abstract Reasoning Test Items*

Model	−2 Log L	χ^2/df	Parameters	Fit
Null	31,382	–	2	–
LLTM	28,768	522.8**	7	.71[1]
2PL-C, cognitive	28,523	40.8**	12	.74[1]
2PL-C, structural	26,152	49.4**	60	.94
2PL	25,406	6.2**	180	1.00

[1] Comparison to 2PL-constrained structural model.
** $p < .01$.

represented different combinations of cognitive variables (Embretson, 1998). The five replicate items for each structure contained different stimuli. The display type was constant within structures but varied across structures. The relationship between the distractors and the key was also equated within structures, as well as between structures, to the extent possible. The key position was randomly assigned.

An empirical tryout of the items supported the cognitive model for generating items with acceptable and similar psychometric properties (Embretson, 1998). Models 1 and 2 both predicted item difficulties to nearly the same level as for APM items; they also predicted the response times. Also like APM, the encoding variables had much less impact on performance than did the rule induction variables. Thus, the construct representation aspect of construct validity was supported by the strong predictions obtained from the cognitive models.

More recently, Model 1 parameters were estimated for a large sample with three replications of thirty item structures (i.e., ninety items) using the 2PL-constrained model (Embretson, 1999), applied with improved estimators (Embretson & Yang, 2002). Table 2 shows the significance and fit for alternative psychometric models of the data. The null model, in which all items are equally difficult and discriminating, is a comparison standard used in the fit index shown in the far right column (see Embretson, 1997a). The goodness of fit statistic divided by its degrees of freedom, χ^2/df, compares successively more complex models for significance increment in fit. It can be seen that the 2PL-constrained model fits more significantly than the LLTM model, which has equal discriminations for all items. Thus, item discrimination parameters increase fit significantly. The structural model, in which a parameter is estimated for each of the thirty structures, fits significantly better than the cognitive model. These results indicate that the cognitive model does not fully reflect differences in the item structures. Finally, the standard 2PL model, where each of the ninety items has unique difficulty and discrimination parameters, fits significantly better than the 2PL structural model, although the increment in fit is not large. These

TABLE 3. *Estimates and Standard Error for Item Difficulty and Item Discrimination for 2PL-Constrained Model*

Feature	Diff. ϕ	σ_ϕ	Slope τ	σ_τ
#Rules	.715*	.031	−.034	.024
Abstract	.784*	.043	−.033	.041
Fusion	−.748*	.078	−.384*	.047
Distortion	−.373*	.052	−.325*	.057
Overlay	.00	.041	−.504*	.038
Constant	−2.142*	.101	1.379*	.087

*$p < .05$.

results suggest that relatively little variability between the items remains after structure is accounted for. Thus, the replicates of the same structure do not vary substantially.

The estimates for the 2PL-constrained model are shown in Table 3. For item difficulty, both number of rules and abstract correspondence, as well as two perceptual variables, are significant predictors of item difficulty. For item discrimination, only the perceptual variables are significant predictors. The negative weights for the variables indicate that fusion, distortion, and overlay are associated with reduced item discrimination.

The difficulties of new items can be predicted from either the structural model or the cognitive model. The 2PL cognitive model yields the following predictions of item difficulty, β', and item discrimination, α':

$$\beta' = -2.142 + .715(\#\text{Rules}) + .784(\text{Abstract}) - .748(\text{Fusion}) - .373(\text{Distortion}).$$

$$\alpha' = 1.379 - .034(\text{Rules}) - .033(\text{Abstract}) - .384(\text{Fusion}) - .325(\text{Distort}) - .504(\text{Overlay}).$$

In the results summarized above, the perceptual variables were not systematically varied within structures. In a recent study, the perceptual variables were varied in an experimental design to examine the strength of their effects (Diehl, 2002; Diehl & Embretson, 2002). Items were generated by crossing eight structures with variations in the perceptual variables. Eight items (with different objects) were created for each structure to observe eight combinations of perceptual features, including display type for multiple cell entries (nested, overlay, adjacent, and platform) and fusion (present versus not present), yielding a total of sixty-four items. Although the perceptual variables had significant impact on item difficulty and response time, again their effect was minor as compared to the rule induction variables. The level of prediction obtained was similar to Embretson (1998), thus yielding further support to the cognitive model.

Item Generation by Artificial Intelligence

Matrix completion items may now be generated from item structures that represent the major cognitive variables, as well as display type. ITEMGEN1 randomly selects stimuli and their attributes to fulfill the structural specifications. All cognitive model variables may be calculated from the structural specifications; hence, item difficulty and discrimination are predictable.

Empirical Tryout of Item Generation

As yet, item generation has not been attempted with the full cognitive approach for the matrix completion items. Further developments to link the generator to a testing system are required, which is expected sometime in 2004.

RELATED APPROACHES TO ITEM DEVELOPMENT

Two other approaches to item development are related to the cognitive design system approach: traditional item development and the item model approach (Bejar et al., 2002; Bejar, 1996). Both of these will be briefly reviewed here.

In the traditional approach, item writing is an art, not a science. Items for intelligence tests are carefully handcrafted by human item writers. Then, the items are submitted for empirical evaluation by calibrating their psychometric properties. Many items do not survive empirical tryout and, consistent with item writing as an art, the reasons for item failure are often unclear. The attrition rate varies substantially for different tests, but rates of 30 to 50% attrition are typical. Surviving items are then calibrated with a psychometric model, particularly IRT models, to be useable for measuring ability. These calibrations are necessary because it is axiomatic to psychometric theory that raw total scores have no meaning because item difficulty levels can vastly influence score levels.

The item model approach (Bejar et al., 2002) is a generative approach, in which existing items are "variablized" to create new items. That is, an item with suitable psychometric qualities serves as a model for new items by allowing one or more of its features to be substituted. For example, an existing mathematics word problem can be variablized by substituting different characters, objects, and settings as well as substituting different numbers. Thus, a family of new items is created. Ability can then be estimated from the new items, without empirical item tryout, as the properties of the item model are assumed inheritable to each new item.

Obviously, the item model approach requires that the item parameters are invariant over the cloned items. Bejar et al. (2002) completed a study of item generation for GRE quantitative items, using the item model

approach. The data strongly support the feasibility of the approach. Analysis of item difficulty and response time indicated a high level of isomorphicity across items within models. Furthermore, the ability estimates from generated items with operational GRE scores were as high as test–retest correlations of two operational GRE tests. Thus, the use of newly generated items has minimal impact on ability estimates.

EVALUATION OF APPROACH: ADVANTAGES AND DISADVANTAGES

The cognitive design system approach to adaptive item generation has several advantages over traditional item development methods. First, new items may be readily developed. Traditional item development procedures do not produce enough items to meet the demands of adaptive testing for large numbers of items. Second, items may be developed to target difficulty levels and adequate psychometric quality. With traditional test development methods, item difficulty levels, at best, can be only informally anticipated. Empirical tryouts typically lead to a high percentage of items rejected for poor quality and inappropriate difficulty. Third, given an adequate calibration of the design principles, new items may be placed in the item bank without empirical tryout. The predicted item parameters from the design variables are sufficient to measure ability. Measurement error increases modestly, but may easily be offset by administering a few more items. Fourth, construct validity is available at the item level. That is, the specific sources of cognitive complexity for each item are given by the weights for the model variables. Fifth, tests may be redesigned to represent specifically targeted sources of item difficulty. The impact of some sources of cognitive complexity can be controlled directly when construct validity is available at the item level. For example, perceptual properties would have minimal impact on solving Abstract Reasoning Test (ART) items if fusion and distortion were eliminated in the items.

The cognitive design system approach has some disadvantages. First and foremost, the approach requires substantial initial effort. Developing a reasonably good cognitive model for an item type requires several empirical studies to support the theory and the models. Whether the approach is practical for a particular test depends on how well the initial cost is compensated for by the unlimited number of new items that can be generated. Second, the approach works best for item types that already have been developed. Although the cognitive design system approach can be applied to new item types, establishing usefulness for measuring individual differences would be required early in the process. That is, the nomothetic span aspect of construct validity should be established by studies on the correlates of scores that are derived from the item type. Nothing in the system prevents applications to new item types, however.

FUTURE

Adaptive item generation may become state of the art for ability and achievement measurement relatively soon. Practically, the increasing need for large numbers of new items makes item generation attractive. Large-scale ability and achievement testing is increasing, not decreasing, and there is special emphasis on repeated measurements. In K–12 education, for example, tests are used increasingly to certify achievement at all levels. In lifespan development, increasing interest in cognitive aging requires longitudinal designs with repeated testing of the same abilities. With the increasing number of tests administered, shorter and more reliable tests are highly desirable. Adaptive testing seems to be the obvious solution. However, adaptive testing requires huge item banks to provide efficient measurement at all levels. Furthermore, as testing becomes more frequent and more important, new items become highly desirable to minimize the response bias that results from previous exposure to items. Item generation by artificial intelligence fulfills these practical needs for new items.

Theoretically, item generation by cognitive design systems also has some advantages. New types of interpretations of test scores are possible when construct validity is available at the item level. When the cognitive sources of item complexity are calibrated in an IRT-based model, ability levels may be described by the processing characteristics of the items appropriate for that level (see Embretson & Reise, 2000, for examples). The continuing debate about the nature of intelligence could take a new direction by referring more specifically to the processes that are involved in performance. The many correlates and relationships of intelligence measurements to other variables may be understood more clearly if the characteristic processing at different ability levels can be explicated.

References

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

Bejar, I. I. (1996). *Generative response modeling: Leveraging the computer as a test delivery medium* (RR-96-13). Princeton, NJ: Educational Testing Service.

Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2002). *A feasibility study of on-the-fly item generation in adaptive testing* (GRE Board Research Report 98-12). Princeton, NJ: Educational Testing Service.

Butterfield, E. C., Nielsen, D., Tangen, K. L., & Richardson, M. B. (1985). Theoretically based psychometric measures of inductive reasoning. In S. E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics* (pp. 77–147). New York: Academic Press.

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.

Carroll, J. B. (1976). Psychometric tasks as cognitive tests: A new structure of intellect. In L. Resnick (Ed.), *The nature of intelligence* (pp. 27–56). Hillsdale, NJ: Erlbaum.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* Cambridge, UK: Cambridge University Press.

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

DiBello, L. V., Stout, W. F., & Roussos, L. (1995). Unified cognitive psychometric assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361–389). Hillsdale, NJ: Erlbaum.

Diehl, K. A. (1998). *Using cognitive theory and item response theory to extract information from wrong responses.* Unpublished Master's thesis, University of Kansas.

Diehl, K. A. (2002). Algorithmic item generation and problem solving strategies in matrix completion problems. Unpublished doctoral dissertation, University of Kansas.

Diehl, K. A., & Embretson, S. E. (2002). Impact of perceptual features on algorithmic item generation for matrix completion problems. Technical Report 02–0100. *Cognitive Measurement Reports.* Lawrence, University of Kansas.

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

Embretson, S. E. (1994). Applications of cognitive design systems to test development. In C. Reynolds (Ed.), *Advances in cognitive assessment: An interdisciplinary perspective* (pp. 107–135). New York: Plenum.

Embretson, S. (1995a). A measurement model for linking individual change to processes and knowledge: Application to mathematical learning. *Journal of Educational Measurement,* vol. 32(3) 275–294.

Embretson, S. E. (1995b). Working memory capacity versus general central processes in intelligence. *Intelligence, 20,* 169–189.

Embretson, S. E. (1997a). Multicomponent latent trait models. In W. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 305–322). New York: Springer-Verlag.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3,* 300–396.

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

Embretson, S. E. (2000). Generating assembling objects items from cognitive specifications. HUMRRO Report No. SubPR98-11. Alexandria, VA: Human Resource Research Organisation.

Embretson, S. E. (2002a). Cognitive models for psychometric properties of GRE quantitative items. Report 02-03 to Assessment Design Center. Princeton, NJ: Educational Testing Service.

Embretson, S. E. (2002b). Generating abstract reasoning items with cognitive theory. In S. Irvine & P. Kyllonen (Eds.), *Item generation for test development* (pp. 219–250). Mahwah, NJ: Erlbaum.

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

Embretson, S. E., & Reise, S. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Embretson, S. E., & Schneider, L. M. (1989). Cognitive component models for psychometric analogies: Conceptual driven versus interactive models. *Learning and Individual Differences,* vol. 1(2), 155–178.

Embretson, S. E., & Wetzel, D. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement, 11,* 175–193.

Embretson, S. E., & Yang, X. (2002). Modeling item parameters from cognitive design features using non-linear mixed models. Symposium paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA, April, 2002.

Embretson, S. E., Schneider, L. M., & Roth, D. L. (1985). Multiple processing strategies and the construct validity of verbal reasoning tests. *Journal of Educational Measurement, 23,* 13–32.

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

Gorin, J. (2002). *Cognitive design principles for paragraph comprehension items.* Unpublished doctoral dissertation, University of Kansas.

Gustafsson, J. E. (1988). Hierarchical models of individual differences in cooperative abilities. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 4, pp. 35–71). Hillsdale, NJ: Erlbaum.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Norwell, MA: Kluwer.

Mislevy, R. J., Sheehan, K. M., & Wingersky, M. (1993). How to equate tests with little or no data. *Journal of Educational Measurement, 30,* 55–76.

Psychological Data Corp. (2002). ITEMGEN1, Item generator for non-verbal intelligence test items. Lawrence, KS: Psychological Data Corporation.

Raven, J. C., Court, J. H., & Raven, J. (1992). *Manual for Raven's Progressive Matrices and Vocabulary Scale.* San Antonio, TX: Psychological Corporation.

Sternberg, R. J. (1977). *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities.* Hillsdale, NJ: Erlbaum.

Whitely,[1] S. E., & Schneider, L. M. (1981). Information structure on geometric analogies: A test theory approach. *Applied Psychological Measurement, 5,* 383–397.

[1] Susan Embretson has also published under the name Susan E. Whitely.