

Developing an Essentially Unidimensional Test With Cognitively Designed Items

Damon U. Bryant

*International Business Machines (IBM)
Dallas, TX*

William Wooten

*University of Central Florida
Department of Psychology*

The purpose of this study was to demonstrate how cognitive and measurement principles can be integrated to create an essentially unidimensional test. Two studies were conducted. In Study 1, test questions were created by using the feature integration theory of attention to develop a cognitive model of performance and then manipulating complexity factors within the model. It was hypothesized that the complexity factors predict item difficulty. Results indicated that some complexity factors predicted difficulty in a relatively small sample. In Study 2, items developed using the cognitive model were integrated with items measuring another factor to create a multidimensional test of spatial reasoning. Results were replicated in Study 2 with a sample of 460 participants. The test met the assumption of essential unidimensionality according to DIMTEST, was moderately correlated ($r = .64$) with the Bennett Mechanical Comprehension Test, and showed little evidence of differential item functioning. Implications are discussed.

Key words: essentially, unidimensional, multidimensional, cognitive design, logistic, IRT

There have been advances in integrating cognitive psychology and psychometrics in creating cognitive ability tests. With developments in the United States (DiBello,

Stout, & Rousso, 1995; Embretson, 1998; Mislevy, 1995), Canada (Gierl, 1997; Gierl, Leighton, & Hunka, 2000), Holland (van der Linden & Hambleton, 1997), and the United Kingdom (Irvine & Kyllonen, 2002), the generation of cognitively designed test items has become a topic of major international interest. This development has been recognized by some as the cognitive design system approach (Embretson, 1995, 2002). First, the goals of measurement are defined and a cognitive or information-processing model of performance is developed. This facilitates identification of complexity factors that are proposed to influence psychometric properties of items (e.g., item difficulty). The model and derived factors can then be applied in manufacturing vast numbers of items with relatively little effort involved by the test developer. Items are then evaluated with respect to the complexity factors. If the model is successful in predicting item characteristics, the manipulation of complexity factors may systematically influence item difficulty and thus have the potential to increase the degree to which the construct is measured and is related to other relevant variables of interest (Embretson, 1983).

There are several advantages in using a cognitive design approach. It takes a more structural view of test development, being concerned about not only functional relationships between constructs (Cronbach & Meehl, 1955) but also the internal or cognitive components that underlie the why and how of the constructs being measured. This is unlike traditional psychometric approaches in that cognitive design methods are able to separate the process from the domain-specific content of the constructs (Kyllonen, 1996). Moreover, if designed on the basis of sound cognitive theories, test items will have a strong theoretical foundation. Unlike traditional test development procedures, this approach, being rooted in cognitive psychology, has the potential to be effectively employed in relatively small samples, which are often encountered in basic and applied human factors or industrial and organizational research.

The focus of this article is on the use of cognitive principles and psychometric approaches in developing cognitive ability tests for use in selection, placement, or training. In this article, two studies are presented. The first study is an example of using cognitive design principles with a relatively small sample. The second study integrates findings from the first study in developing an essentially unidimensional test, which measures one dominant dimension.

STUDY 1: COGNITIVE DESIGN

Many cognitive psychologists argue that intelligence is nothing more than working memory capacity (Kyllonen, 1996; Kyllonen & Christal, 1990). Others posit that intelligence is a function of two distinct types of attentional processing, such as focused versus broad or serial versus parallel processing (Messick, 1996). These two views are not necessarily inconsistent with each other. Baddeley (1992, 1993) sug-

gested that of the three systems of working memory (i.e., the phonological loop, the visuo-spatial sketch pad, and the central executive), the one most responsible for attention is the central executive, which integrates and coordinates activities in memory. Selection, evaluation, and integration of relevant information can be seen as important precursors for skill acquisition and performance (Baddeley, 1993). This investigation will demonstrate the use of a cognitive theory that involves these processes while minimizing the importance of domain-specific knowledge via simple sensory discrimination.

The pioneers in intelligence research have sought to establish a link between intelligence and pure sensory discrimination. Sir Francis Galton first hypothesized that a positive relation between sensory discrimination and intelligence exists, but evidence to support his claim was anecdotal (Deary, 2000). Spearman, establishing a relation between measures of academic performance and sensory discrimination, has been more successful with a mean correlation between sensory discrimination and school grades of .39 (Deary, 2000). This evidence suggests that the same cognitive processes involved in lower-level, sensory functioning may generalize to higher-level, cognitive processing. Researchers have also speculated that cognitive theories of sensory functions may provide insight into general intellectual capacity (Messick, 1996).

Messick (1996) suggested that the way in which information is processed in vision is the same process used in memory to scan, select, and integrate information. One theory that provides a means by which this proposition can be evaluated is the feature-integration theory of attention (Treisman & Glade, 1980; Treisman & Gormican, 1988). The theory posits that objects in the visual field are processed in one of two ways: parallel processing, which involves divided attention and simultaneous intake of information, and serial processing, which requires focused attention on one object in the visual field. Akin to a spotlight that shines on an object in a dark room, attention in vision is narrowed or focused depending on what is being processed. If target objects and nontarget objects differ in more than one relevant dimension, then they are processed in serial with focused attention, which involves selecting a particular feature within a dimension and using working memory to conjoin different features of the objects to form a coherent whole. This is feature integration or visual binding of objects in their spatial locations (Treisman, 1999). After the whole object is compared to other nontargets in the visual field, a choice is made about the correct target.

With respect to scanning in long-term memory, a target to be retrieved may be similar to nontargets on one or more dimensions and different along other dimensions in memory. The selection of the correct target may be a function of retrieving and integrating declarative and procedural knowledge within the domain-specific area. To the extent that targets require integration of more than one feature or component of declarative and procedural knowledge, then the difficulty in conjoining the target by the central executive in memory is increased because of the increased load on working memory capacity.

Using a definition of working memory (Baddeley, 1986) and factor analytic methods, Kyllonen and Christal (1990) created several tests of working memory and found that on other domain-specific tests of cognitive ability (e.g., verbal, spatial, and quantitative), the best predictor of test performance is working memory. There is also support for generalizing their findings. Kyllonen (1996) reviewed research, showing working memory as being the best predictor of declarative and procedural learning in training. It is also the best predictor of transfer in two different interventions, computer programming and logic gate. Moreover, knowledge of working memory and lower-level cognitive processing may provide a basis for developing measures of higher-level processing using the cognitive design systems approach.

First, this study employs the cognitive design approach in (a) developing an ability test that uses information from lower-level, sensory processing and working memory to understand higher-level, cognitive processing and (b) evaluating the test's psychometric properties. For the purpose of this study, a cognitive theory is chosen that is purported to resemble the cognitive processes involved in the selection, evaluation, and integration of knowledge (Messick, 1996). We develop a model of performance based on theory and derive factors that have the potential to influence difficulty.

Cognitive Theory Selection

Although Neisser (1963) uncovered general principles involving types of visual processing, making the distinction between parallel (divided attention) and serial processing (focused attention), he noted that there were large individual differences in the degree to which persons could make discriminations in objects. Today it is generally accepted that when a stimulus is multidimensional in the context of other multidimensional stimuli that vary along more than one dimension, visual processing is serial and self-terminating. Moreover, when objects are similar to each other in their component features, errors in selecting targets from nontargets are likely to occur. This investigation attempts to develop a test to measure individual differences in the ability to make visual discriminations. The theoretical framework guiding the development of this measure is the feature-integration theory of attention (Treisman & Glade, 1980).

Abundant evidence suggests certain primitive dimensions of objects are detected in the preattentive stages of attention, requiring little conscious effort (Treisman & Gormican, 1988). These dimensions include size, color, closure (shape), curvature, orientation, and feature presence (Treisman, Sykes, & Glade, 1977). In other words, when objects differ in only one of the aforementioned dimensions, the relevant features within that dimension "pop out." For example, within the dimension of shape, a triangle that is embedded within the context of circles will pop out when participants are prompted to identify the presence of a triangle. However, when objects differ along more than one dimension (e.g., color and shape), targets and nontargets be-

come difficult to discriminate. A blue triangle embedded within the context of red circles, blue circles, and red triangles will require serial processing and focused attention when participants are prompted to identify the presence of a blue triangle. At this point, serial processing and focused attention become necessary to integrate features and distinguish targets from nontargets.

Model Development and Selection of Complexity Factors

For the purpose of this study, we create a model of performance using the following dimensions derived from feature-integration theory of attention: color, shape and lines, orientation, and feature presence. Dimensions or complexity factors are preattentive in that they are detected in parallel and are separated before conscious awareness. These dimensions (e.g., color) have varying features (e.g., red, green or yellow) conjoined only by serial processing and focused attention on the spatial location within the object (Treisman, 1977). When judgments are made about the similarity or dissimilarity of a target, a multistage process occurs. First, objects are encoded and decomposed. Discriminable features (squares and circles) are detected within a dimension (shape). A relevant dimension (e.g., color) is chosen in which features within that dimension (e.g., red or blue) are evaluated across objects in a serial manner. If a feature is detected that differs across all objects, then the search process terminates, and the object with the feature that differs from all others is selected. If a feature is not detected within this dimension, this process is repeated for the next dimension until it is realized that features within objects must be conjoined to detect a target from nontargets. Accordingly, features of an object are then conjoined by visually processing their spatial locations, and the objects are now compared on the basis of their conjunction or binding of features. The conjunction of features of an object must be kept in working memory long enough to compare it to each object that subsequently becomes the focus of attention. This model of performance just described is presented in Figure 1. It should be noted that within this model both declarative and procedural knowledge are enclosed within long-term memory and any model of performance must include these components (Kyllonen, 1996). Because declarative and procedural knowledge are domain-specific, we don't explicitly extend these subcomponents of long-term memory in our model due to the focus on process.

Several errors can occur in making visual discriminations. First, features of objects may not be conjoined together because attention was not focused or processing was not serial. Second, because of demands on working memory capacity, a correct target feature within a dimension may differ from features in other objects and may not be detected as being relevant because of a large number of features within a dimension. Third, features within a target object may not be conjoined correctly with other features in their respective spatial locations due to numerous features within and across dimensions of an object, thus limiting the amount that

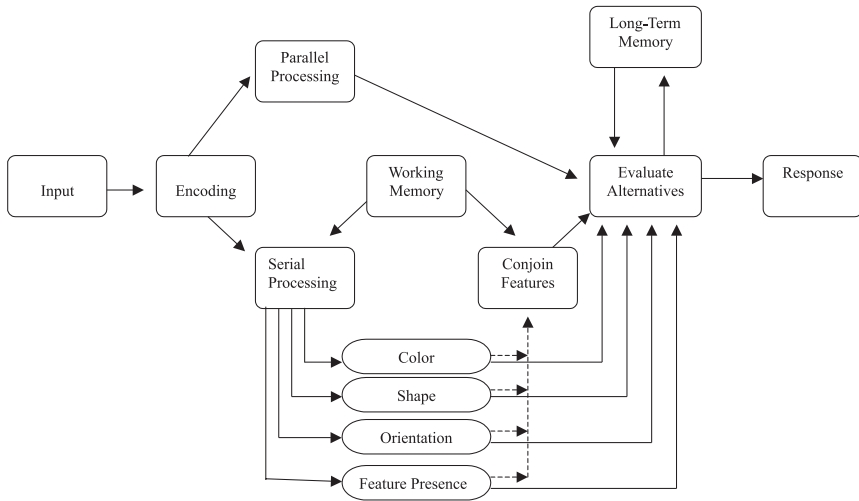


FIGURE 1 General model of cognitive performance for object discrimination task.

can be simultaneously processed and conjoined. It is apparent that the putative cause of most of the errors is the limit on working memory capacity. Thus, based on the literature of feature integration theory, the following hypotheses are made when participants are required to discriminate between multidimensional objects:

- H1: As the number of similar features within the dimension of color increases in multidimensional targets (correct answers) and nontargets (incorrect answers), the difficulty of items increases.
- H2: As the number of similar features within the dimension of shape and lines increases in multidimensional targets and nontargets, the difficulty of items increases.
- H3: As the number of different orientations of objects within items increases, the difficulty of items increases.
- H4: When correct responses require participants to detect the presence of a feature within a correct target as opposed to the absence of a feature within an object, the item difficulty will decrease.

Methods

Participants

Eighty-two participants were sampled from a large metropolitan area in the southeastern United States. The sample had equal numbers of males and females, the

average age of the sample was 24.77 ($SD = 5.59$), and the average number of years of education was 14.58 ($SD = 1.66$). The majority of the sample was Caucasian (80%).

Materials

Object discrimination test. Forty items were developed using the cognitive model derived from feature-integration theory of attention (Treisman & Glade, 1980; Treisman, Sykes, & Glade, 1977). Color, shape, orientation, and feature presence were manipulated so there would be a large variety of features within each dimension (i.e., different colors, different shapes, and different orientations). Because these dimensions are easily recognized in everyday vision in normal individuals, the domain-specific content should not limit item performance and thus provide a better understanding of the process. Base items with eight choices were created. Then, each of the dimensions was independently manipulated to create the 40 items. There was only one correct answer for each item. The participants were asked to select the object that was different from all of the rest. A sample item is shown in the participant instructions in the Appendix. The person's total score on the measure, which is computed as a linear combination of correct responses, is a measure of the person's ability. Scores are standardized with a mean of zero and unit standard deviation.

Procedures

Participants were told that they could withdraw participation for whatever reason to ensure their involvement was voluntary. Participants read the general overview of the study that detailed the procedures to be followed. If participants agreed to continue, they were asked to sign an informed consent. Each person was then given one scantron sheet, an instruction page, and the measure of object discrimination. Pretesting was done with a subsample to determine an adequate time to complete test items. After providing varying time limits to persons within the subsample, 25 min was the time judged most adequate to complete the test. Thus, the remaining subsample (62 participants) was allotted approximately 25 min to complete the test. The results from this subsample are presented. After completion of the test, the participants were debriefed and thanked for participating.

Analyses

The analyses of the data were conducted in two phases that involved different units of analysis. In the first phase, the psychometric properties of the items were assessed; the unit of analysis is the person. Reliability estimate of scores on the test was determined by computing Kuder-Richardson formula 20. Then, correlates with external variables were determined. Systematic variation in scores was a necessary condition for continuing to Phase II.

In the Phase II, Hypotheses 1 through 4 were evaluated. Consistent with other investigations using this approach (Embretson, 2002), the unit of analysis was the item. The independent variables were complexity factors derived from the cognitive model. Each item was coded according to the dimensions of shape and line similarity, color similarity, feature presence, and object orientation. All of the independent variables were transformed to normal scores with a mean of zero and unit standard deviation.

The difficulty of items was the dependent variable. Difficulty in this study was operationally defined as the location on the ability scale where a person will have a 50% chance of getting an item correct (Hambleton, Swaminathan, & Rogers, 1991). This value was obtained from NOHARM (Fraser & McDonald, 1988) in exploratory mode specifying a one-dimensional solution. Prior research showed that NOHARM parameter estimates are stable with as few as 100 participants (Maydeu-Olivares, 2001). To account for the possibility of unstable results with small samples, the natural logarithm of the odds of getting the item correct for each item is also used as a measure of difficulty and is compared to the NOHARM results.

Results

Phase I

Means, standard deviations, and correlations of the demographic variables and the object discrimination test (ODT) are presented in Table 1. The reliability estimate of scores for this measure was .85. There were small correlations among scores and demographic variables. In this particular sample, there was a relationship between sex and ODT scores, $r = .27, p \leq .05$. This relationship remained statistically significant after controlling for years of education, $pr = .27, p \leq .05$. Females ($M = 27.41, SD = 5.51$) scored higher than males ($M = 23.86, SD = 6.95$) on this measure, $t(60) = -2.71, p \leq .05$. There was also a small and significant relationship between age and ODT scores, $r = -.28, p \leq .05$.

TABLE 1
Means, Standard Deviations, and Intercorrelations Among Object
Discrimination Test and Demographic Variables

<i>Variables</i>	<i>M</i>	<i>SD</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>
1. Object Discrimination Test	24.4	6.76	—			
2. Age	28.1	5.27	-.29*	—		
3. Education	14.1	2.45	.20	-.33*	—	
4. Sex	1.32	0.48	.27*	-.21	.28*	—

Note. $N = 62$. For sex: male = 0, female = 1.

* $p \leq .05$.

Phase II

Means, standard deviations, and correlations of complexity factors and difficulty are shown in Table 2. Regression was used to test Hypotheses 1 through 4. The dependent variable was item difficulty and the independent variables were shape/line similarity, color similarity, object orientation, and feature presence of each of items. When using the $\ln(\text{odds})$ of a correct response as the dependent variable, the overall model was significant, $R = .63$, $F(4, 35) = 5.71$, $p < .001$, accounting for approximately 40% of the variance. Similar results were found when using the difficulty parameter estimated using NOHARM, $R = .62$, $F(4, 35) = 5.70$, $p < .001$, accounting for approximately 39% of the variance. See Table 3 for regression coefficients and significance values.

Evidence from two different measures of difficulty produced comparable results. Hypotheses 2 and 3 were supported. Hypothesis 2 stated that as the number

TABLE 2
Means, Standard Deviations, and Intercorrelations of Difficulty and Complexity Factors

Variables	<i>M</i>	<i>SD</i>	1	2	3	4	5
1. Difficulty	-0.48	1.16	—				
2. Color	3.31	1.38	.57**	—			
3. Shape	5.13	3.46	.53**	.43**	—		
4. Orientation	1.77	1.09	-.05	-.36*	-.48**	—	
5. Feature presence	0.28	0.46	-.13	-.31	.06	.08	—

Note. $N = 40$.

* $p < .05$. ** $p < .01$.

TABLE 3
Regression Model Predicting Difficulty and $\ln(\text{Odds})$ From Complexity Factors

Variables	<i>Ln(Odds)</i>			<i>Difficulty (b)</i>		
	<i>B</i>	<i>SE B</i>	β	<i>B</i>	<i>SE B</i>	β
Constant	.72	.16		-.82	.22	
Color	-.33	.19	-.22**	.52	.27	.31**
Shape	-.68	.20	-.55***	.87	.27	.52***
Orientation	-.55	.19	-.44**	.68	.26	.41**
Feature presence	.21	.17	.17	-.29	.23	-.17
Multiple <i>R</i>		.63***			.62***	
R^2		.40			.39	
Adjusted R^2		.33			.33	

Note. $N = 40$. All independent variables are in z score form.

* $p < .10$. ** $p < .05$. *** $p < .005$.

of shapes and lines increases in multidimensional targets and nontargets, difficulty of item increases. The regression coefficients for shapes and lines across both analyses were statistically significant below the .005 level (see Figure 2). Using the regression equation predicting difficulty (b_i) estimated by NOHARM in Table 3, we graphed the influence of the complexity factor of shapes and lines on item difficulty. As the number of shapes or lines increased in multidimensional targets and nontargets, the item difficulty increased. Hypothesis 3 stated that as the number of orientations of targets increases in multidimensional targets and nontargets, difficulty of item increases. The regression coefficients across both analyses were statistically significant. The item response functions were similar to those in Figure 2 and are not presented. There was no support for Hypotheses 1 and 4.

Discussion

The results suggest that the cognitive design system approach was partially successful in developing a test with adequate psychometric properties. The approach was also useful in deriving complexity factors that can be used to create numerous test items. In Phase I, the test was also found to have an adequate reliability estimate at .85, which is above the recommended value of .70 suggested by Nunnally and Bernstein (1994).

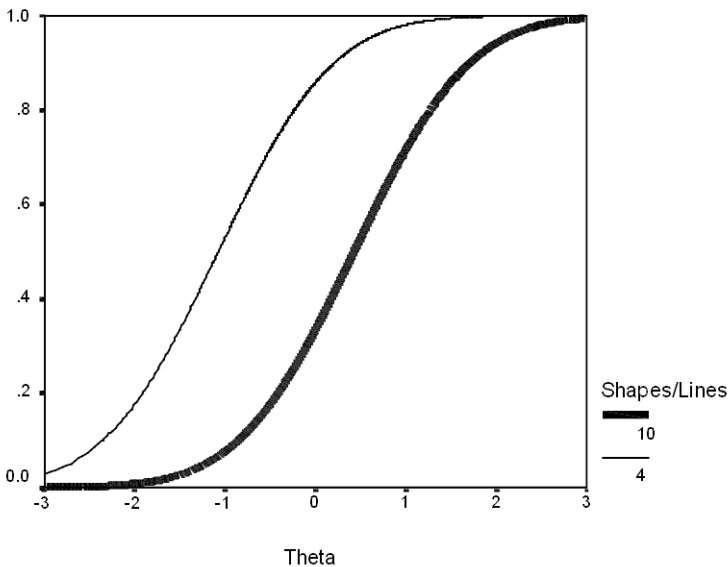


FIGURE 2 Predicted item difficulty as a function of shape.

In Phase II of this study, approximately 40% of the variance in item difficulty was explained by increasing working memory capacity on the dimensions of shape and orientation. The effects of these dimensions on item difficulty were meaningfully predicted by the feature-integration theory of attention. As the number of common orientations and shapes among targets (correct answers) and nontargets (incorrect answers) increased, the difficulty of the items increased. Because features of targets (correct responses) and nontarget (incorrect answers) were similar but differed in their spatial orientation, errors could have occurred if numerous features within or across dimensions were increased. Moreover, the greater the number of orientations of the items, the greater working memory capacity needed to respond correctly. This may provide insight into separating aspects of domain-specific knowledge into dimensions that can be meaningfully manipulated, which may have utility in job or task analysis or training interventions.

Feature presence was not a significant predictor of difficulty. A possible explanation for this null result can be attributed to the level of manipulation of this variable. Because there were a few levels of features within this dimension, the manipulation may not have been adequate. All of the other dimensions had five or more levels. Future research may consider this null result and increase the variability within this dimension. This may also have relevance to domain-specific knowledge. Perhaps some dimensions within a content area may or may not be critical dimensions that influence the difficulty of acquiring knowledge or skill. Future research could investigate this further.

The evidence in this investigation suggests that the complexity factors of shape and lines and orientation derived from feature-integration theory of attention were successful in influencing the psychometric property of item difficulty. The results in this study are consistent with other research employing a cognitive design system approach. Using a matrix processing theory, Embretson (2002) found that two complexity factors, number of rules and abstract correspondence accounted for 59% of the variance in performance on a test of matrix completion (i.e., Advanced Progressive Matrices). This investigation is unique in that results were found with a relatively small sample. Although sample size can be seen as a limitation, confidence can be placed in the findings due to strong theoretical and empirical foundations of feature integration theory. This approach may have utility in other fields of psychology, such as human factors or industrial and organizational psychology, where large samples are not always possible. Moreover, organizations could perhaps capitalize on the strengths of this small sample test design using cognitive principles.

This investigation has several limitations. First, with a relatively a small sample size, it is possible the results are unstable. The sample size has an influence on the estimation of the metric used as a proxy of item difficulty. Notwithstanding research that has successfully used NOHARM in samples as small as 100 (Maydeu-Olivares, 2001), it is typically employed as large sample technique. In this study, it is vulnerable in producing a rather unstable parameter estimate, which

would perhaps lead to biased results. Future research may need to conduct a study in several phases. In the first phase, items are developed from the model and evaluated on a relatively small sample of approximately 50, as in this investigation. This provides preliminary information on what complexity factors are successful in altering difficulty of items. In the second phase, the results from the preliminary investigation are used to manufacture more items based on findings in phase one; these items are evaluated on a larger sample of about 500. At this point, programs such as NOHARM and BILOG will be able to produce more stable item characteristics and ability estimates. In the third phase, the items are investigated for the presence of differential item functioning.

Another threat to validity is the dimensionality of object discrimination. Because the items were developed by manipulating several components in the cognitive model, the test may be multidimensional, which will require more than one score being reported. With the small sample size, the dimensionality cannot adequately be assessed. Although the assumption of dimensionality has been relaxed considerably to that of essential dimensionality (Stout, 1990), it is uncertain that this test is essentially unidimensional or multidimensional. Dimensionality assessment has important implications for the assessment of differential item functioning (Mazor, Hambleton, & Clauser, 1998). Incorporating an assessment of dimensionality in Phases 2 and 3 of the aforementioned test construction process may ameliorate this problem. DIMTEST procedures (Nandakumar & Stout, 1993) may be used to test this assumption.

Notwithstanding the limitations, this investigation demonstrates that cognitive psychology and psychometrics can be efficiently used to guide the test development process by employing a cognitive design systems approach. This is just an example of how a measure can be designed for purposes of diagnosing the propensity for effectively acquiring knowledge and skills. Moreover, it provides a foundation for the investigation of general cognitive processes while separating domain-specific content to be trained.

STUDY 2: DEVELOPING AN ESSENTIALLY UNIDIMENSIONAL TEST

The evidence in Study 1 suggests the cognitive design approach can be implemented in small samples and provide guidance in the development of complex items. In addition, recommendations were made for a program of research in which developing cognitively based items, assessing dimensionality, and investigating differential item functioning are primary issues of concern before collecting other validity-related evidence. Following the recommendation in Study 1, we manufactured items in this study similar to those in Study 1 except that color was not manipulated. These items based on feature integration theory of attention were

also combined into a larger test of work skills (Prien, Wooten, & Prien, 2000) with the goal of creating a test with one dominant dimension from items that measure different dimensions. Accordingly, multidimensional parameter estimates from NOHARM and MIRT statistics are used to develop an essentially unidimensional measure of work skills using methods described by Reckase, Ackerman, and Carlson (1988).

Research has shown that unidimensional tests can be developed from items that measure one or more than one latent trait by computing measurement angles based on parameter estimates from multidimensional item response theory models (Reckase, Ackerman, & Carlson, 1988). Items are then assembled on the basis of their measurement angles from the k dimensions. In the two-dimensional case, items are selected to create a fan-like structure. One set of items is selected to measure the first dimension (e.g., 0° to 30° from Dimension 1), a second set of items is chosen to measure the second dimension (e.g., 0° to 30° from Dimension 2), and a third set measures both dimensions with measurement angles ranging from 30° to 60° on either dimension. Then, the dimensionality of the test is evaluated using one of a variety of procedures, such as Yen's (1984) Q_3 -statistic or DIMTEST (Nandakumar & Stout, 1993).

For the purpose of this investigation, DIMTEST procedures are employed to evaluate the assumption of essential unidimensionality. This conditional covariance-based approach evaluates the null hypothesis that there is one dominant dimension, $H_0: d = 1$, against the alternative hypothesis that the number of dimensions is greater than one, $H_0: d > 1$. This was done by factor analyzing tetrachoric correlation coefficients, creating two assessment tests and one partitioning test based on factor analysis or expert judgment, and calculating a T-statistic. This procedure has proven to be useful in simulated and nonsimulated research (see Nandakumar & Stout, 1993) and is used in this investigation to determine if a test developed from multidimensional items is "essentially unidimensional" (Stout, 1990).

Most item response theory models assume monotonicity, local independence, and unidimensionality (Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1991). The latter assumption is difficult to satisfy because items used to develop most tests are multiply determined; that is, more than one latent trait is needed to produce a correct response (Reckase, 1985; Reckase & McKinley, 1991). Consequently, the assumption requiring unidimensionality has been relaxed to one of "essential unidimensionality" (Stout, 1990), which allows for the presence of minor dimensions so long as one dominant dimension is present. When the assumption of essential unidimensionality is violated, the utility of models based on one score is questionable and has important implications in the identification of potentially biased items (Ackerman, 1992; Roussos & Stout, 1996).

Several statistics based on parameter estimates within multidimensional item response theory (MIRT) have been developed to gain a better understanding of the complexity of items that involve the use of more than one ability for a correct re-

sponse (Reckase, 1985; Reckase & McKinley, 1991). These statistics have been used primarily with the multidimensional 2-parameter logistic model (M2-PL), which is expressed as follows:

$$P(y_{ij} = 1 \mid \mathbf{a}_i, d_i, \theta_j) = [1 + \text{Exp}(-L)]^{-1}, \quad (1)$$

where y_{ij} is the response on item i by person j , \mathbf{a}_i is a vector of k discrimination parameters for item i , $[a_{1i}, a_{2i}, \dots, a_{ki}]'$, k is the number of dimensions, θ_j is a vector of k ability parameters for person j , $[\theta_{1j}, \theta_{2j}, \dots, \theta_{kj}]'$, and d_i is a scalar representing the intercept ($-\sum a_{ki}b_{ki}$) and is related to difficulty (Reckase, 1985), $L = D(a_i', \theta_j + d_i)$, D is equal to a scaling constant 1.7 or 1.

Because the linking functions (i.e., normal and logistic) of item response theory models have the same parameter structure (Embretson & Reise, 2000; McDonald, 1982), multidimensional discrimination (MDISC), multidimensional item difficulty (MDIFF), polar coordinates of items in the θ space, and angles of the item vectors from the orthogonal axes of the abilities can be computed using parameter estimates from multidimensional item response theory models (for a description of the theoretical framework, see Reckase & McKinley, 1991).

The point of steepest slope in the ability space is known as multidimensional discrimination,

$$\text{MDISC}_i = \|\mathbf{a}_i\| = (a_i', a_i)^{1/2}, \quad (2)$$

where $\|\mathbf{a}_i\|$ represents the length of vector \mathbf{a}_i that is computed as the square root of the sum of squared elements of vector \mathbf{a}_i . MDISC_i is interpreted in the same manner as the discrimination parameter (a_i) in unidimensional item response theory (IRT). The difficulty of the item is the signed distance from the origin of the multidimensional space to the point of steepest slope. Reckase (1985, 1997) conceptualized multidimensional item difficulty, MDIFF , as the signed distance from the origin of the latent ability space to the point of the steepest slope of the item i . The formula for multidimensional difficulty is given by

$$\text{MDIFF}_i = -d_i(\|\mathbf{a}_i\|)^{-1} = -d_i/\text{MDISC}_i, \quad (3)$$

and it is interpreted in the same way as the difficulty parameter (b_i) in unidimensional IRT. Reckase (1985) has shown MDIFF_i to be equal to the unidimensional measure of difficulty (b_i) when there is only one dimension.

The angle, α_{ki} , of the item vector from the axis orthogonal to dimension θ_k is computed as

$$\alpha_{ki} = \arcsin(a_{ki} \text{MDISC}_i^{-1}), \quad (4)$$

where α_{ki} is the discrimination parameter of item i on dimension k . Items that are representative of the ability being measured will have small angles from the orthogonal axis of θ_k , and items that are not representative of θ_k will have larger angles relative to others in the ability space. In a two-dimensional space when items

are positively related to both abilities, angles range from 0° to 90° and sum to 90° . Polar coordinates in a two-dimensional space are computed by multiplying the cosines of each of the angles by $MDIFF_i$. These polar coordinates are plotted in the first and third quadrants in a Cartesian plane, representing monotonically increasing relationships of items to the dimensions being measured. These statistics have been used primarily in the development of essentially unidimensional tests in the United States (Reckase, Ackerman, & Carlson, 1988).

Recently, researchers have been applying MIRT statistics to investigate differential item functioning (Bryant, Wooten, Forde, & Reynolds, 2003; Mazor, Hambleton, & Clauser, 1998; Oshima & Miller, 1992). It has been shown in multidimensional tests that certain items exhibit differential item functioning (DIF) when there are differences among examinee groups in conditional distributions of ability given a fixed level of a second, relevant ability. To circumvent this potential problem, researchers cluster items according to a validity sector (Ackerman, 1994) for each of the k -dimensions, using MIRT statistics such as the angle of the item vector from the axis orthogonal to the intended dimension or ability. Then, a procedure such as logistic regression (Zumbo, 1999) is easily employed for the detection of uniform and nonuniform DIF. However, MIRT statistics have not been applied beyond the multidimensional two-parameter logistic model. Moreover, little has been done in the way of integrating cognitive design principles and DIF within a multidimensional framework. This investigation attempts to amalgamate several areas of research in developing an essentially unidimensional test.

In Study 1, we demonstrated the use of the cognitive design approach in creating test items for one of the k dimensions in an essentially unidimensional test. In this study, MIRT statistics are used to develop an essentially unidimensional measure of spatial reasoning. Then, DIMTEST (Nandakumar & Stout, 1993) procedures are employed to evaluate the assumption of essential unidimensionality. MIRT statistics and parameter estimates from NOHARM (Fraser & McDonald, 1988) are used to determine validity sectors in evaluating DIF with logistic regression. Finally, some evidence of validity with a measure closely related to the intended construct is presented.

Method

Participants and Materials

Participants ($N = 460$) answered 65 items that were developed to measure a variety of work skills (Prien et al., 2000). This sample was recruited from a university setting in the southeastern United States. The 65 items were designed to measure the ability to understand and apply mechanical relationships in practical situations. Some items were designed in a manner consistent with the model developed in Study 1. Approximately 60% of the sample was female.

Procedures

The objective is to determine the dimensionality of this 65-item measure and design a test that will meet the assumption of essential unidimensionality. First, factor analysis of tetrachoric correlations will provide some evidence of dimensionality. This will be followed by a more formal test of the data, DIMTEST. If the data are multidimensional, then MIRT parameter estimates from NOHARM will be used to compute multidimensional discrimination (MDISC), the signed distance to the point of steepest slope (MDIFF_{*i*}), and angles of the item vectors from the orthogonal axes of the *k*-dimensions. A version of the test is then assembled to create a fan-like structure using item measurement angles. The dimensionality is reevaluated and a DIF analysis is conducted. We conclude by correlating scores on this measure with that of a well-established measure of mechanical ability (Bennett, 1969).

Results

Evidence of Internal Structure: Dimensionality Assessment

Because the 65 items involved content from several different components, including 10 items developed from feature integration theory, we expected more than one factor to be present. The “fac” program within DIMTEST was used. This program computed tetrachoric correlations and then performed a principle-axis factoring of the $i \times i$ matrix. This was done to determine the eigenvalues for the first three factors of the data and gave a coarse indication of the number of dimensions. Eigenvalues for the first three orthogonal factors were 11.13, 3.62, and 2.81, which indicates that the data are not strictly unidimensional. A formal evaluation of the data was then conducted. The items for the assessment tests (1 and 2) and the partitioning subtest were chosen using the automatic selection procedure used by DIMTEST. Both the conservative (T_c) and more powerful test (T_p) rejected the null hypothesis that there is only one dominant dimension, $T_c = 7.62, p < .01$ and $T_p = 8.44, p < .01$. The eigenvalues and formal tests provided sufficient evidence supporting the notion that the 65 items are not essentially unidimensional. Accordingly, multidimensional item response theory parameter estimates from NOHARM were used in conjunction with MIRT statistics (see Equations 2, 3, and 4) to better understand the complexity of these items in creating an essentially unidimensional test.

Test Construction

NOHARM in exploratory mode estimated parameters for the multidimensional model with two dimensions using all 460 participants. Although *k* dimensions can be requested in exploratory mode, a simple two-dimensional case is illustrated. Parameter estimates (a and d_i) were obtained from the matrix of

coefficients for thetas, which is a matrix ($i \times k$) of discrimination parameters, and the final $f(0)$ vector, which is a vector ($i \times 1$) of intercept terms analogous to d_i in Equation 1. The root mean square residual (RMSR) of sample item covariances to item covariances reproduced by the parameters has been advocated by researchers as a measure of item fit (Embretson & Reise, 2000; Hattie, 1985). The reproduced matrix of item covariance provided a reasonable fit to the data. The RMSR was .011.

After estimating parameters and computing MIRT statistics, criteria for item inclusion were specified. First, items were selected based on their discrimination (MDISC) and item angles. For a two dimensional case, all item discrimination parameters must be related to dimensions in a monotonically increasing manner, so items that did not meet the assumption of monotonicity on one or both dimensions were excluded. Twenty-five items failed to meet this criterion. The remaining items had measurement angles that were between 0° and 90° , and the sum of the angles for both dimensions was 90° . In addition, a reasonable estimate of MDISC was necessary. The MDISC value chosen was .50 or higher. A total of 20 items remained that met the aforementioned criteria for inclusion. The parameter estimates and MIRT statistics are listed in Table 4.

The validity sector (Ackerman, 1992) for each of the two dimensions was chosen and ranged from 0° to 35° . Items measuring Dimension 1 (θ_1) had vectors with angles of 0° to 35° from the orthogonal axis of Dimension 1 (see column α_1 in Table 4). Items measuring Dimension 2 (θ_2) had item vectors with angles of 0° to 35° in column α_2 of Table 4. A third set of items appeared to measure both dimensions; as such, items had vectors with angles between 36° and 64° from the axes of both dimensions θ_1 and θ_2 . Next, the content of the items was evaluated to understand what the test was measuring.

Evidence Based on Content

Dimension 1. Based on the measurement angle α_1 , the first dimension consisted of eight items. Six of these items were constructed using the cognitive design approach. Ten items of this type were created for this study. Only shape and orientation were manipulated. Consistent with the results in Study 1, an analysis of the 10 cognitively based items indicated that item difficulty was once again predicted by the complexity factors, $R = .82$, $F(2, 7) = 7.15$, $p < .05$, accounting for 67% of the variance in item difficulty, expected R^2 was .22. Also consistent with the results in Study 1, females ($M = 6.77$, $SD = 1.6$) scored slightly higher than males ($M = 6.19$, $SD = 1.95$) on these items, $t(458) = -3.12$, $p < .005$. It should be noted that items used in Study 1 were not used in this investigation. Both sets were designed using the same cognitive principles based on feature integration theory of attention. In light of this fact, results were comparable across studies.

TABLE 4
Multidimensional Item Response Theory Statistics Based on NOHARM
Parameter Estimates

Item	a_{i1}	a_{i2}	d_i	MID			
				α_1	α_2	MDIFF _{<i>i</i>}	MDISC _{<i>i</i>}
1	0.52	0.09	0.73	10	80	-1.38	0.53
2	0.52	0.13	0.46	14	76	-0.87	0.54
3	0.91	0.31	0.95	19	71	-1.01	0.95
4	0.77	0.34	1.24	23	67	-1.47	0.84
5	0.89	0.44	0.91	24	66	-0.94	0.97
6	0.52	0.28	1.51	29	61	-2.63	0.57
7	0.46	0.29	1.34	32	58	-2.45	0.55
8	0.67	0.43	0.58	33	57	-0.73	0.80
9	0.41	0.35	1.82	41	49	-3.39	0.54
10	0.76	0.68	1.47	42	48	-1.44	1.02
11	1.92	2.19	-0.42	49	41	0.14	2.91
12	0.34	0.46	-0.31	53	37	0.53	0.57
13	0.64	0.92	-0.09	55	35	0.08	1.12
14	0.92	1.32	-0.41	55	35	0.25	1.61
15	0.42	0.67	-0.49	58	32	0.62	0.79
16	0.76	1.21	-0.26	58	32	0.18	1.42
17	0.39	0.65	-0.36	59	31	0.48	0.76
18	0.59	1.04	-0.31	61	29	0.26	1.20
19	0.61	1.16	-0.54	62	28	0.41	1.31
20	0.43	0.90	-0.53	65	25	0.53	1.03

Note. MDIFF = Multidimensional Item Difficulty; MDISC = Multidimensional Item Discrimination; a_{ik} = discrimination of item i on dimension k ; α_k = angle from orthogonal axis of dimension k ; and d_i = intercept term for item i .

Dimension 2. Based on the measurement angle α_2 , the second dimension consisted of eight items. These items involved the use of spatial and planning skills. Each examinee was given a map with parallel, horizontal, and diagonal lines connecting different locations represented by square blocks. Certain travel constraints were placed on the examinee in determining which routes in the map would consume the least amount of time (e.g., a diagonal route consists of 90 miles, rate of speed is 60 miles per hour). Questions involved different constraints and different locations. This skill required examinees to select the best route among several alternatives by integrating information such as speed, time constraints, and distance to travel. There was no significant difference between males and females on this dimension.

There were four items measuring both dimensions with angles between 36° and 64° on either dimension. Two of these items were developed using the cognitive design approach described in Study 1, and the other two were items based

on spatial reasoning. This two-dimensional measure designed to be essentially unidimensional was evaluated using DIMTEST procedures.

Assessment of Dimensionality

Tetrachoric correlations were computed as in the initial evaluation of dimensionality. Eigenvalues for the first three factors in this 20-item measure were 9.63, 2.25, and 1.37, accounting for 48%, 11%, and 7% of the variance, respectively. This is in sharp contrast to the variance accounted for by the first three extracted factors in the original 65 items (i.e., 17%, 6%, and 4%). Items for the assessment tests (1 and 2) and the partitioning subtest were chosen using the automatic selection procedure used by DIMTEST. The conservative (T_c) and more powerful test (T_p) failed to reject the null hypothesis that there is one dominant dimension, $T_c = 0.16, p < .05$ and $T_p = 0.14, p < .05$. This evidence supports the view that the 20-item measure is essentially unidimensional despite clear information that the test contained two dimensions, one created by the cognitive design approach involving feature integration and another involving spatial reasoning/planning skills.

For comparison purposes, MIRT statistics of difficulty ($MDIFF_i$) and discrimination ($MDISC_i$) were compared to one-dimensional parameter estimates of difficulty (b_i) and discrimination (a_i) for the essentially unidimensional test. The NOHARM estimate of b_i and $MDIFF_i$ were positively correlated, $r = .99, p < .001$. Multidimensional discrimination ($MDISC$) and a_i parameter estimate were also positively related, $r = .98, p < .001$.

Evidence of Internal Structure: Differential Item Functioning

Because dimensionality has important implications in the detection of potentially biased items, we conducted a DIF analysis using logistic regression. We performed analyses consistent with those outlined by Swaminathan and Rogers (1990) using terms to represent ability (i.e., standardized total score, θ_i), group membership (g , an effects coded variable sex), and the group by ability interaction ($g \times \theta_i$) in the logit function. We also employed logistic regression with two ability estimates based on the validity sectors computed in this study (see Table 4). There have been studies that investigated the degree to which logistic regression is able to detect DIF in tests that are multidimensional (Bryant et al., 2003; Mazor et al., 1998; Oshima & Miller, 1992). Bryant et al. (2003) used DIMTEST procedures and validity sectors in the detection of uniform and nonuniform DIF, while Mazor et al. (1998) were successful in detecting DIF using NOHARM factor loadings.

After an effects coded variable for group membership (males vs. female) was created, standard scores on each of the two abilities, interaction terms of group membership and each of the abilities were computed. These variables were entered into a logistic regression analysis for the detection of uniform DIF, which is a sig-

nificant difference in item difficulty between groups after controlling for ability estimates, and nonuniform DIF, which is a significant difference in item discrimination between groups after controlling for ability (Mellenberg, 1982). In this case, two interaction terms were created, one for detection of differences in discrimination on the first ability ($g \times \theta_1$), and the other for the detection of differences in discrimination on the second ability ($g \times \theta_2$).

Results indicated that when an essentially unidimensional model (i.e., with one ability estimate) was used, items showed no DIF at $p \leq .01$ level of significance. However, when $p \leq .05$ was the standard of evidence, one item (19) showed nonuniform DIF as indicated by a significant likelihood ratio test for the group by ability interaction, $\chi^2(1) = 5.71, p < .05$. This item is illustrated in Figure 3.

When the multidimensional DIF detection model was used (i.e., with two ability estimates determined from validity sectors of 0° to 35°), there were no items that showed DIF at the $p \leq .01$ level of significance. However, three items (3, 13, and 19) showed significant DIF at the .05 level. Item 19 was consistently detected across the two analyses. All of the detected items showed nonuniform DIF. Moreover, all items in the multidimensional analysis exhibited nonuniform DIF only on the first dimension, which consisted primarily of items developed using the cognitive design approach. Although the explanation of DIF is not the primary focus of this study, interested readers

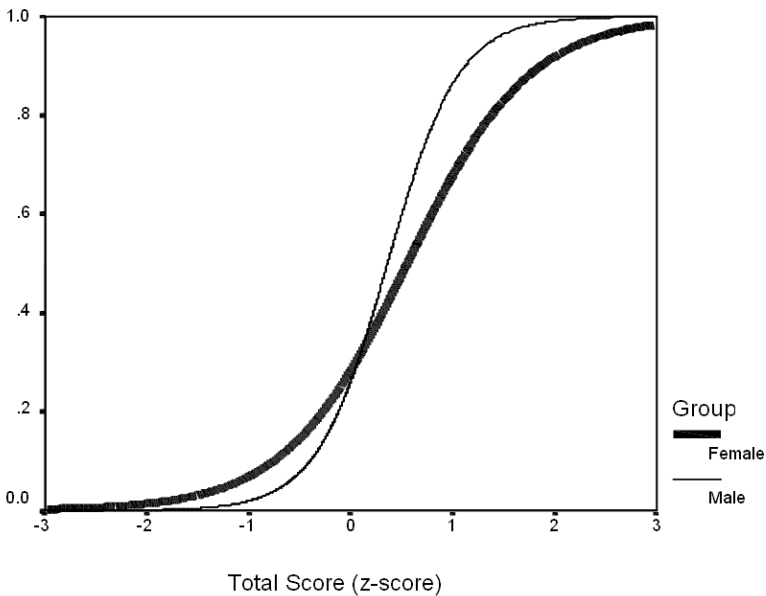


FIGURE 3 Item 19 exhibiting nonuniform DIF as a function of sex.

are referred to Roussos and Stout (1996) for potential reasons. Future research could perhaps explore the extent to which multiple complexity factors contribute to DIF.

Evidence of External Relations

Of the 460 participants in this study, a subsample of 50 also answered questions to a 68-item version of the *Bennett Mechanical Comprehension Test* (BMCT; Bennett, 1969). This measure has been used as a selection test in employment and has demonstrated evidence of job-related validity as high as .38 (Muchinsky, 1993). The relation between the BMCT and the 20-item measure of work skills is significant, $r = .64$, $p < .01$, sharing approximately 40% common variance. This evidence supports the view that feature integration skills and spatial reasoning are associated with mechanical comprehension.

GENERAL DISCUSSION

The primary purpose of this study was to demonstrate how cognitive and measurement principles can be integrated within a multidimensional framework to create an essentially unidimensional test. Using feature integration theory of attention and a cognitive model of performance, we developed several items by manipulating complexity factors and were successful in predicting item characteristics. Then, using the same principles in Study 2, we were able to replicate the results in a larger sample. The majority of the items developed in the second study using the cognitive design approach (80%) were psychometrically sound to withstand the criteria used to develop the essentially unidimensional test. Once the essentially unidimensional test was developed and evaluated for dimensionality using DIMTEST, it showed respectable validity with a currently existing measure of mechanical aptitude. This investigation proved to be a promising start to integrate several areas of research.

In terms of the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999), several requirements of the test have been met. We have provided evidence on the internal structure of the test, elucidated cognitive processes involved in item performance, investigated the measure for the presence of DIF, and presented some validity evidence with an external measure. It is hoped that a better understanding of the measurement process is achieved by incorporating cognitive and psychometric principles in the test development process.

ACKNOWLEDGMENT

A previous version of this article was presented at the 17th annual conference of the Society for Industrial and Organizational Psychology, Toronto, Ontario, Canada.

REFERENCES

- Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*, 67–91.
- Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education, 7*, 255–278.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Baddeley, A. (1986). *Working memory*. New York: Oxford University Press.
- Baddeley, A. (1992). Working memory. *Science, 255*, 556–559
- Baddeley, A. (1993). Working memory or working attention. In A. Baddeley & L. Weiskrantz (Eds.), *Attention: Selection, awareness and control* (pp. 152–170). New York: Oxford University Press.
- Bennett, G. K. (1969). *Manual for the Bennett Mechanical Comprehension Test*. Cleveland, OH: The Psychological Corporation.
- Bryant, D. U., Wooten, W., Forde, D., & Reynolds, A. M. (April, 2003). *Separating benign and adverse differential item functioning: A simulation study*. Paper presented at the 18th annual conference for the Society of Industrial and Organizational Psychology, Orlando, FL.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.
- Deary, I. J. (2000). Simple information processing and intelligence. In R. J. Sternberg (Ed.), *Handbook of intelligence* (pp. 267–284). New York: Cambridge University Press.
- DiBello, L., Stout, W., & Roussos, L. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361–389). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179–197.
- Embretson, S. E. (1995). Developments toward a cognitive design system for psychological tests. In D. Lubinski & R. V. Dawis (Eds.), *Assessing individual differences in human behavior: New concepts, methods, and findings* (pp. 17–48). Palo Alto, CA: Davies-Black Publishing.
- 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. (2002). Generating abstract reasoning items with cognitive theory. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 219–250). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares factor analysis. *Multivariate Behavioral Research, 23*, 267–269.
- Gierl, M. J. (1997). Comparing the cognitive representations of test developers and students on a mathematics achievement test using Bloom's taxonomy. *Journal of Educational Research, 91*, 26–32.
- Gierl, M. J., Leighton, J. P., & Hunka, S. (2000). Exploring the logic of Tatsuoka's rule-space model for test development and analysis. *Educational Measurement: Issues and Practice, 19*, 34–44.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury, CA: Sage.
- Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*, 139–164.
- Irvine, S. H., & Kyllonen, P. C. (Eds.). (2002). *Item generation for test development*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

- Kyllonen, P. C. (1996). Is working memory capacity Spearman's g ? In I. Dennis & P. Tapsfield (Eds.), *Human abilities: Their nature and measurement* (pp. 49–75). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Kyllonen, P. C., & Christal, R. E. (1989). Cognitive modeling of learning abilities: A status report of LAMP. In R. F. Dillion, & J. W. Pellegrino (Eds.), *Testing: Theoretical and applied issues* (pp. 146–173). New York: Praeger.
- Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working memory capacity? *Intelligence*, *14*, 389–433.
- Maydeu-Olivares, A. (2001). Multidimensional item response theory modeling of binary data: Large sample properties of NOHARM estimates. *Journal of Educational and Behavioral Statistics*, *26*, 51–71.
- Mayer, R. E., & Sims, V. K. (1994). For whom is a picture worth a thousand words? Extensions of a dual-coding theory of multimedia learning. *Journal of Educational Psychology*, *86*, 389–401.
- Mazor, K. M., Kanjee, A., & Clauser, B. E. (1995). Using logistic regression and the Mantel-Haenszel with multiple ability estimates to detect differential item functioning. *Journal of Educational Measurement*, *32*, 131–144.
- Mazor, K. M., Hambleton, R. K., & Clauser, B. E. (1998). Multidimensional DIF analysis: The effects of matching on unidimensional subtest scores. *Applied Psychological Measurement*, *22*, 357–367.
- McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement*, *6*, 379–396.
- Mellenberg, G. J. (1982). Contingency table models for assessing bias. *Journal of Educational Statistics*, *7*, 105–118.
- Messick, S. (1996). Human abilities and modes of attention: The issue of stylistic consistencies in cognition. In I. Dennis & P. Tapsfield (Eds.), *Human abilities: Their nature and measurement* (pp. 77–96). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Mislevy, R. J. (1995). Probability-based inference in cognitive diagnosis. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 43–71). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Muchinsky, P. M. (1993). Validation of intelligence and mechanical aptitude tests in selecting employees for manufacturing jobs. *Journal of Business and Psychology*, *7*, 373–382.
- Nandakumar, R., & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics*, *18*, 41–68.
- Neisser, U. (1963). Decision time without reaction time: Experiments in visual scanning. *American Journal of Psychology*, *76*, 376–385.
- Nunnally, J. C., & Bernstein, I. C. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
- Oshima, T. C., & Miller, M. D. (1992). Multidimensionality and item bias in item response theory. *Applied Psychological Measurement*, *16*, 237–248.
- Prien, E. P., Wooten, W., & Prien, K. O. (2000). *The work skills test: A gender bias free method of measuring mechanical comprehension*. Unpublished manuscript, University of Central Florida.
- Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, *9*, 401–412.
- Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271–286). New York: Springer-Verlag.
- Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement*, *25*, 193–203.
- Reckase, M. D., & McKinley (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement*, *15*, 361–373.
- Roussos, L., & Stout, W. (1996). DIF from the multidimensional perspective. *Applied Psychological Measurement*, *20*, 335–371.
- Stout, W. F. (1990). A new item response theory modeling approach with application to unidimensional assessment and ability estimation. *Psychometrika*, *55*, 293–326.

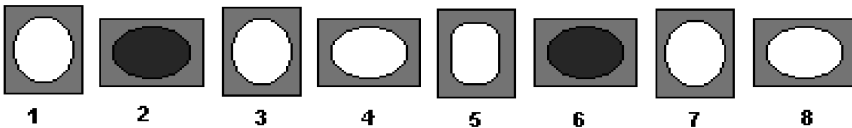
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27, 361–370.
- Treisman, A. (1977). Focused attention in the perception and retrieval of multidimensional stimuli. *Perception and Psychophysics*, 22, 1–11.
- Treisman, A. (1999). Feature binding, attention, and object perception. In G. W. Humpreys, J. Duncan, & A. Treisman (Eds.), *Attention, space, and action: Studies in cognitive neuroscience* (pp. 91–111). New York: Oxford University Press.
- Treisman, A., & Gormican, S. (1988). Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review*, 95, 15–45.
- Treisman, A., & Souther, J. (1985). Search asymmetry: A diagnostic for preattentive processing of separable features. *Journal of Experimental Psychology: General*, 114, 285–309.
- Treisman, A. M., & Glade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12, 97–136.
- Treisman, A., Sykes, M., & Glade, G. (1977). Selective attention and stimulus integration. In S. Dornic (Ed.), *Attention and performance VI* (pp. 333–361). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.
- Yen, W. M. (1984). Effects of local independence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8, 125–146.
- Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (dif): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, Ontario, Canada: Directorate of Human Resources Research and Evaluation: Department of National Defense.

APPENDIX

Participant Instructions

You will be presented with 40 questions. Each question has eight objects that are similar to each other. Your task is to select the object among the eight that is different from all the rest. You may base your answer on one (e.g., line color) or more than one aspect of the object (e.g., color and shape). An example item is provided below.

- 1) Select the object that is different from all the rest.



The correct answer is **5**. Objects 1, 3, and 7 are the same. Objects 4 and 8 are the same, and Objects 2 and 6 are the same.

Please complete the questions as accurately as possible. The time allowed to answer the items is approximately 25 minutes, after which your answer sheet will be collected by the researcher. You may begin once the researcher tells you. When you are finished, raise your hand and the researcher will collect your information. If you have any questions about what you will be doing, let the researcher know.