Journal of Educational Measurement
Spring 1992, Vol. 29, No. 1, pp. 25–50

Measuring and Validating Cognitive Modifiability as an Ability: A Study in the Spatial Domain

Susan E. Embretson
University of Kansas

Measuring cognitive modifiability from the responsiveness of an individual's performance to intervention has long been viewed (e.g., Dearborne, 1921) as an alternative to traditional (static) ability measurement. Currently, dynamic testing, in which cues or instruction are presented with ability test items, is a popular method for assessing cognitive modifiability. Despite the long-standing interest, however, little data exists to support the validity of cognitive modifiability measures in any ability domain. Several special methodological difficulties have limited validity studies, including psychometric problems in measuring modifiability (i.e., as change), lack of appropriate validation criteria, and difficulty in linking modifiability to cognitive theory. In this article, relatively new developments for solving the validation problems are applied to measuring and validating spatial modifiability. Criterion-related validity for predicting learning in an applied knowledge domain, as well as construct validity, is supported.

Cognitive modifiability long has been deemed an important aspect of ability because it is a measure of change in performance as a function of learning. Dearborne (1921), at the first conference on measuring intelligence, complained that static tests measure what the examinee already knows, not necessarily what he or she can learn. Currently, cognitive modifiability is assessed by dynamic testing procedures in which ability test items are presented with cues or instruction to determine the effect of intervention on an individual's performance. Dynamic tests have appeal over static (traditional) tests in many educational settings because they seem more instructively relevant. Bransford, Delclos, Vye, Burns, and Hasselbring (1987) suggest that dynamic tests can provide one or more of the following: (a) information about learning processes rather than learning products, (b) assessments of the responsiveness of an examinee to instruction, and (c) diagnostic information about effective intervention techniques.

The validity of dynamic tests, particularly in comparison to traditional static tests, is crucial to establishing cognitive modifiability as meaningful. In the Lidz (1987) edited volume on dynamic testing, several chapters present results relevant to construct validity. That is, data are given on issues such as convergent and discriminative validity with respect to static intelligence tests and other traits (e.g., Budoff, 1987; Rand & Kaniel, 1987), transfer of the dynamic training effects to similar tasks (Campione & Brown, 1987), group differences in modifiability (e.g., Tzuriel & Klein, 1987), impact on teacher's perceptions of examinee's competencies, and the perceived usefulness of

dynamic test information for instruction (e.g., Vye, Burns, Delclos, & Bransford, 1987). Such data provide a promising start toward construct validity but certainly do not fully develop the theoretical meaning of cognitive modifiability.

Noticeable by their absence, unfortunately, are results that support cognitive modifiability measures as useful for predicting learning. The issue is particularly important because the negative findings of earlier studies are still cited. For example, Woodrow (1938, 1946) failed to support modifiability as either a general learning ability or as related to educational learning. Stake's (1961) later study on learning ability yielded somewhat more positive results, in that individual differences in modifiability (measured as a curvature parameter estimate in a learning model) correlated with school achievement. However, incremental validity for the modifiabilities over traditional static aptitude tests was not supported.

The few available contemporary studies on the incremental validity of dynamic tests provide only weak support, due to both the strength of the findings and the relatively small sample sizes ($Ns < 100$). For example, Carlson and Weidl (1979) found that performance on the Raven's Coloured Progressive Matrices correlated more highly with teachers' ratings of achievement when administered under their most elaborate dynamic testing procedure than under the standard (static) instructions. However, the difference between the regressions was not tested, and the extremely small sample sizes limit generalizations in any case. Another type of validity study involves pretest to posttest comparisons. If the pretest is a traditional static test, then incremental validity is supported when the posttest correlates more highly with learning than the pretest. Babad and Budoff (1974), for example, find that the posttest scores correlate more highly with various indexes of achievement than pretest scores for low-IQ subjects. However, the magnitude of the difference is quite small ($r = .29$ vs. .35), and the regressions are not tested for differences. Further, small sample sizes again limit the findings. Other examples are available in the German psychological research (see Guthke, 1982), but, unfortunately, apparently they have not been translated into English.

Several factors contribute to the sparse data on criterion-related validity. First, current dynamic tests do not employ optimal procedures for measuring individual differences in modifiability. Studies which employ scores on a test that is administered after intervention (e.g., Babad & Budoff, 1974; Carlson & Weidl, 1979) have used an index in which initial level is confounded with modifiability. An alternative index is a change score, which is the simple or the residualized difference between a (static) pretest and a posttest that follows the intervention. However, the paradoxes surrounding the interpretability of simple change scores (see Bereiter, 1963), which include the negative relationship of change score reliability to test-retest reliability, the dissimilar meaning of change at different levels of initial performance, and the spurious negative correlation with initial ability, have led to the fact that some psychometricians have recommended abandoning their use whenever possible (Cronbach & Furby, 1970).

Second, appropriate criteria for validating dynamic tests are not easily obtained. Ideally, because dynamic tests are assumed to be better predictors of learning, they should be validated against achievement measures that clearly reflect learning and the acquisition of competence. As noted by Glaser and Bassok (1989), the merging of cognitive psychology into instructional psychology can guide the development of theories of instruction and learning, which in turn can guide the assessment of learning and competence. However, readily available indicators of learning, such as school grades and teachers' ratings, typically are not based on a theory of instruction and learning. In fact, typical achievement measures may reflect prior or prerequisite knowledge rather than the change in knowledge states that is associated with instruction. Further, the impossibility of precisely monitoring an individual's learning process in a classroom setting, as well as the subjective element in grading, further confounds school grades as measures of learning. Problems with validation criteria for new ability tests have, in part, prompted major research programs on measuring individual differences in the acquisition of applied knowledge domains (e.g., Kyllonen & Christal, in press; Kyllonen & Shute, in press).

Third, establishing the construct validity of a cognitive modifiability measure requires elaborating construct representation as well as nomothetic span (see Embretson, 1983, for definitions). The nomothetic span of a cognitive modifiability measure can be elaborated by relating modifiability to other measures of individual differences, particularly standing on a learning criterion or other traits. If cognitive modifiability shows different correlations than initial ability with other measures of individual differences, then supporting the nomothetic span for a cognitive modifiability measure implies that the dynamic testing procedure has changed the nature of performance. However, nomothetic span studies do not indicate directly how the dynamic testing procedure influences the cognitive processes, strategies, or knowledge structures that are involved in solving test items. Thus, establishing the construct validity of a particular measure of modifiability also requires understanding construct representation. It should be noted that construct representation depends on the specific dynamic testing procedures that are employed. Even for the same items, different interventions or cues probably will have varying influences on construct representation.

Several relatively recent developments are applied to measuring and validating a cognitive modifiability measure in the current study. To resolve some psychometric difficulties with change measurement, individual differences in modifiability, as well as initial ability, are estimated in a new psychometric model, the multidimensional Rasch model for learning and change (MRMLC; Embretson, 1991a). This model will be described briefly, below. To resolve some difficulties with validation criteria, learning is measured by precise indexes of performance in the acquisition of an applied knowledge domain for which the subjects have little prior knowledge. Finally, changes in construct representation are examined by comparing mathematical models of item response time and accuracy between the pretest and the posttests.

Cognitive Modifiability in the Spatial Domain

A previous study (Embretson, 1987) examined the validity of both initial spatial ability and modifiability to predict learning in an applied knowledge domain (i.e., computer operations). Substantial increases in performance were observed on the spatial ability test after intervention. Mathematical modeling of accuracy on the pretest and posttest also indicated that construct representation had changed as well. In this study, the learning criterion was measured precisely as accuracy and response time from a behavior sample on a criterion task for which subjects had little advance knowledge. Both initial spatial ability and spatial modifiability had significant independent contributions to predicting learning.

The current study extends the Embretson (1987) study in several ways. First, an appropriate psychometric model for estimating individual differences in modifiability was available (Embretson, 1991a). Second, stronger support for a spatial modifiability measure was possible, as a much larger sample size was planned. Third, modifiability was measured on a spatial ability task that is more clearly connected to cognitive theory. That is, a spatial ability test was available in which the item stimulus features were designed to vary the difficulty of specific spatial processes according to an empirically supported cognitive model of the task (Embretson & Waxman, 1989). Fourth, the changes in cognitive processing associated with dynamic testing could be tested more adequately because the collection of item response times was planned to allow mathematical modeling of processing durations. Fifth, two modifiability measures were available, rather than just one. That is, one modifiability measured the impact of cue training on performance (as in Embretson, 1987) while the other modifiability measured the impact of strategy training on the cognitive model that was used to specify the test items. Sixth, a learning criterion was available with properties similar to the Embretson (1987) criterion but with the additional advantage of providing information about the process of learning. That is, performance was measured at the various stages of learning.

Dynamic testing was hypothesized to influence both the construct representation and the nomothetic span aspects of construct validity (see Embretson, 1983). Because dynamic testing provides instruction or cues that are designed to change cognitive processes, structures, or strategies, it was hypothesized that the construct representation of the spatial ability test would change after intervention. In turn, the nomothetic span of the spatial ability test was also expected to change. Specifically, the changes in construct representation were expected to influence the validity of the test for predicting learning in the applied knowledge domain.

Method

Dynamic test instrument. The Spatial Learning Ability Test (SLAT; Embretson, 1989) was developed to measure spatial ability from a spatial folding task, which is administered in a multiple choice format. Figure 1 presents three sample items. The examinee's task is to select the alternative that can be made by folding the stem downward. The three distractors are views of folded stems

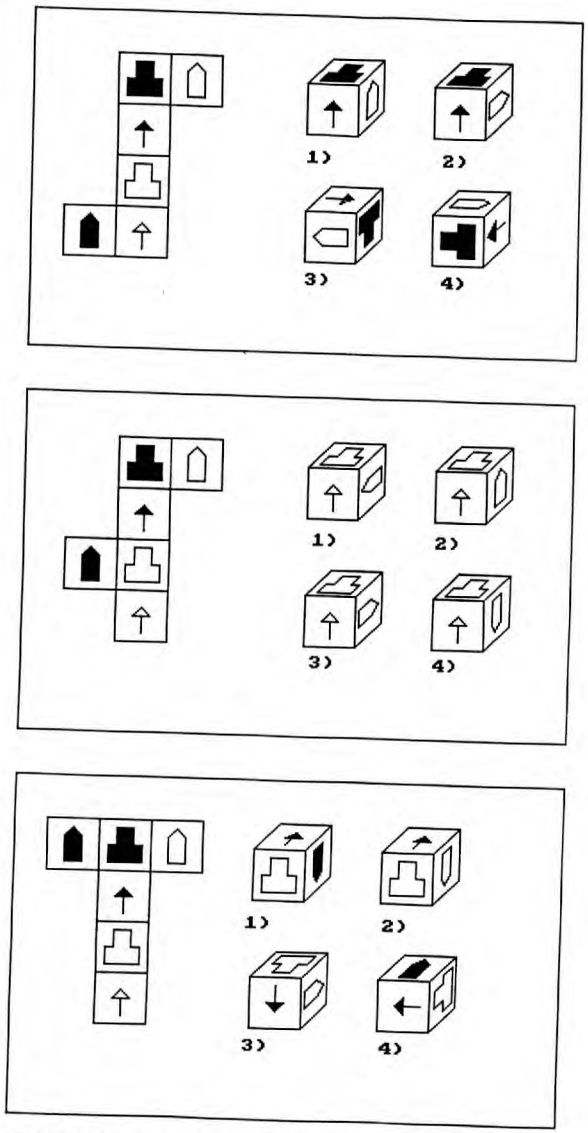

FIGURE 1. *Three items from the SLAT*

that have different configurations of the side markings than the configuration shown in the stem.

The items on the test were selected to represent various combinations of features that influence processing difficulty. Embretson and Waxman (1989) developed an attached folding model to explain processing on the spatial folding task. As typical for processing models of choice tasks, encoding and decision processes were postulated. However, the inherently spatial processes in the model are attaching and folding. In attaching, the unfolded stem is superimposed mentally on the folded alternative by aligning the markings on

two sides. Then, in folding, the third side is folded mentally. Two features of stimuli can be manipulated to influence the difficulty of attaching and folding: the degrees of rotation required to attach the unfolded stem and the number of surfaces carried to fold the stem to the alternative, respectively. In the top item in Figure 1, the correct answer (Option 2) requires zero degrees of rotation to attach the stem as does Option 1, a distractor. However, Option 4 requires a 90° rotation while Option 3 requires a 180° degree rotation. All alternatives in this item require only one surface to be carried for folding. In the middle item, all alternatives require zero degrees rotation, but three surfaces must be carried during folding. Notice also that the distractor set varies between the top and middle item. That is, in the middle item, all alternatives have the same orientation while in the top item each distractor requires different degrees of rotation. Last, the bottom item can be solved without actually folding the stem. Subjects can solve the item by noting that the third side appears to the right of the adjacent markings in both the stem and the correct answer (Option 2).

SLAT consists of three structurally equivalent test forms. Each form contains 18 spatial items, which are postulated to be solved by fully folding the stem, and 6 position items (like the bottom item on Figure 1), which are postulated to be solvable without the folding process. The position items are included to detect possible changes in strategies over occasions. The 18 spatial items were generated by crossing the number of surfaces carried (1, 2, or 3 surfaces) with the degrees of rotation (0°, 90°, or 180°) and the type of distractor set (uniform or mixed) in a 3 × 3 × 2 design. The 6 position items were generated by crossing degrees of rotation with distractor type in a 3 × 2 design. All position problems had two surfaces to be carried. Additionally, four linking items appear on a form when it is administered as a pretest.

Dynamic test procedure. The three forms of the SLAT were computer-administered within a 1-hour session, under three conditions. First, a SLAT form was administered as a pretest, using standard static test instructions and a maximum exposure rate of 60 seconds per item. Second, a SLAT form was administered as a posttest, following a short (about 10 minutes) intervention in which 15 items were presented with cues. The cues were the physical analogue of the stems, three wooden cutouts of cubes with hinged sides. The items were presented on the CRT in blocks of five to accompany each wooden cutout. Subjects controlled their own exposure rate. Third, a SLAT form was administered as a second posttest, following a brief strategy training unit on the attached folding model. Both response time and accuracy on each item were recorded, using the MICROCAT (Vale, 1984) item presentation program.

Estimating modifiability. Embretson (1991a) developed a multidimensional Rasch model for measuring learning and change (MRMLC). With MRMLC, one or more modifiabilities may be estimated, as well as initial ability, from repeated measurements of ability. MRMLC is a multivariate generalization of the simple Rasch model in which a simplex structure is postulated to govern the involvement of initial ability and the modifiabilities in performance on two or more occasions. Thus, for a pretest, only initial ability is involved in performance. At the second or later measurement, performance depends on initial

ability as well, but, because some condition precedes the later measurements (e.g., targeted cues, instruction, practice, stress, or just time), performance also depends on one or more modifiabilities.

MRMLC may be written as follows:

$$P(x_{ij} = 1) = \frac{\exp\left(\sum_{m=1}^{k} \theta_{jm} - b_i\right)}{1 + \exp\left(\sum_{m=1}^{k} \theta_{jm} - b_i\right)}, \quad (1)$$

where θ_{j1} is initial ability at the first occasion, $\theta_{j2} \ldots \theta_{jM}$ are modifiabilities that correspond to the second and later occasions, and b_i is item difficulty.

Embretson (1991a, 1991b) shows how MRMLC resolves two of Bereiter's (1963) three paradoxes in measuring modifiability by a change score and resolves one aspect of the third paradox. First, the reliability paradox for change scores disappears. The paradox is that change scores are most reliable when the pretest to posttest correlation is low, which indicates low test reliability if both tests measure the same trait. Placing modifiability in a multidimensional model resolves the paradox because a lower pretest to posttest correlation is predictable from the involvement of additional dimensions (i.e., modifiability) in the posttest performance. Second, the varying meaning of raw change scores at different initial ability levels is directly accommodated by MRMLC. Embretson (1991a, 1991b) shows that, because MRMLC is an item response theory model, different raw changes are expected for persons with equivalent modifiabilities when their initial abilities differ. Third, the spurious negative element in the correlation of change scores with initial ability is partially reduced in applications of MRMLC by removing scaling artifacts that contribute to a spurious negative element in raw gain scores.

The primary advantage in using MRMLC estimates is for individual measurement. As in other item response theory models, latent abilities are estimated from patterns of item responses, rather than linearly derived from total score. MRMLC not only provides more optimal scaling of initial ability, as unidimensional Rasch models, but also provides modifiability estimates that are corrected for differences in initial level. The many advantages of item response theory models in interpreting and comparing abilities, item selection, equating, and adaptive testing apply also to MRMLC.

For criterion-related validity studies, the main advantage in using MRMLC estimates is to provide correlations and regression coefficients that are directly interpretable as the impact of modifiability, which is the target measure from a dynamic testing procedure. However, for fixed content tests, MRMLC estimates probably would not yield substantially higher levels of predictive validity than pretest and posttest scores. As for unidimensional Rasch item models, MRMLC initial ability estimates correlate very highly with raw pretest score (r = .994), and MRMLC modifiability correlates somewhat less highly (r = .953) with raw gain (see Embretson, 1991b). However, higher levels of predictive

validity could be expected for MRMLC estimates that are derived from adaptive tests, particularly due to improved precision in estimating modifiability at the extreme levels.

Identifying the parameters of MRMLC requires multiple groups so that each item may be observed under each condition. A practical method for accomplishing this is to counterbalance test forms across occasions with Latin square designs. In the current study, three groups were required, because three occasions of measurement (a pretest and two posttests) were to be observed. The parameters for MRMLC may be estimated by programs for logistic item response models in which item parameters may be constrained across groups. Embretson (1991a) derives conditional maximum likelihood estimators for the item parameters and maximum likelihood estimators for the multidimensional ability parameters. Basically, estimating the item parameters involves conditioning ability on each occasion within each group. Thus, for the current study, nine technical groups are required. Estimates are obtained from a single calibration in which parameter estimates are linked across (real) groups by common items that are administered under the same condition for all groups (i.e., the four linking items on the pretests) and across occasions by constraining each item's difficulty to a common parameter, b_i, when observed in the three conditions. Conditions vary systematically across groups.

Criterion task. The criterion task was a computerized tutoring program in a technical area for which subjects had little or no advanced knowledge. Thus, performance scores maximally reflect new learning rather than prior knowledge. The tutoring program (see Kyllonen & Stevens, in press) was developed to teach electronics trouble-shooting by presenting schematic diagrams, called *logic gates,* which represent the flow of electrical currents through various points. Training concerned the conditions under which electricity would flow. Both positive conditions and negation conditions were taught. Because cognitive theories of task performance typically assume that negation demands greater processing, henceforth the negation logic gates will be labeled as the complex learning task while the positive logic gates will be labeled as the simple learning task.

Both accuracy and response time were recorded on the learning task. Additionally, the process of learning also was measured, because performance was measured at eight and six successive learning stages, respectively, on the simple and the complex learning tasks.

Subjects. The subjects were 582 Air Force recruits who had completed a computer-administered version of the SLAT. Subjects were assigned randomly to one of three test form order conditions. Subjects with zero or perfect scores on any SLAT form were excluded from the analysis of the accuracy data, due to the impossibility of estimating abilities for these subjects in MRMLC. This exclusion left a sample of 504. The subjects had completed about 3 weeks of basic training at the Lacklund Air Force Base. The tutoring program was completed by a subset of the sample. The simple learning task was completed by 298 subjects while the complex learning task was completed by 162 subjects.

Results

Descriptive statistics. Table 1 presents descriptive statistics for the SLAT and for the MRMLC ability and modifiability estimates. It can be seen that SLAT standard scores are increasing. The MRMLC estimates, θ_{jm}, which are shown in the logit scale, also indicate substantial change. The positive means for both modifiabilities indicate increasing performance levels for most subjects. The standardized effect from the pretest to the second posttest is .83 while the standardized effect from the pretest to the first posttest is .55. Table 1 also shows that response times are decreasing across occasions.

Table 2 presents the correlations of SLAT standard scores and MRMLC estimates. It can be seen that the correlations between SLAT scores decrease as time between the measures increases, as expected for a simplex pattern. In contrast, MRMLC estimates show low intercorrelations, also as expected, because MRMLC estimates reduce negative bias in the correlation between initial ability and modifiability. Furthermore, it can be seen that, although SLAT pretest scores are highly correlated with MRMLC initial ability, SLAT posttest scores have only moderate correlations with the MRMLC modifiabilities, as expected. That is, the posttests measure initial ability as well as one or more modifiabilities.

Table 3 presents descriptive data on the learning task. For the simple learning task, the mean accuracies are increasing while the mean response times are decreasing over training. Furthermore, for both accuracy and response time, the standard deviations are decreasing. For the complex learning task, although accuracy levels are not as high as for the simple learning task, the mean accuracies are also increasing. Further, although the response times are much higher on the complex learning task, they also decrease substantially over training.

Table 1
Descriptive Statistics of SLAT Standard Scores and MRMLC Abilities

Variable	$\bar{X}$	SD
Accuracy - standard scores		
1) Pretest	99.74	14.51
2) Posttest 1	106.54	18.17
3) Posttest 2	110.54	19.49
Accuracy - MRMLC estimates		
4) Initial ability	-.16	.90
5) Modifiability 1	.50	.91
6) Modifiability 2	.25	.98
Response time		
7) Pretest	24.04	7.49
8) Posttest 1	19.56	6.97
9) Posttest 2	15.54	6.10

Table 2
Correlations Between SLAT Standard Scores and MRMLC Abilities

Variable	(1)	(2)	(3)	(4)	(5)	(6)
Standard scores						
1)Pretest	1.00					
2)Posttest 1	.77	1.00				
3)Posttest 2	.69	.79	1.00			
MRMLC Abilities						
4)Initial ability	.98	.75	.68	1.00		
5)Modifiability 1	.06	.61	.38	.04	1.00	
6)Modifiability 2	.02	-.13	.42	.02	-.25	1.00

Construct Representation

Construct representation was studied by mathematically modeling subjects' response times and accuracies on the 72 nonlinking items from the processes that are postulated to underlie the spatial folding task. Because two types of items were contained in the SLAT, a contrast to compare spatial with position items was scored. The remaining effects were operationalized separately for the two types of SLAT items. For the spatial items, the difficulty of the attaching and folding processes, as described above, was operationalized by

Table 3
Descriptive Statistics for Learning in an Applied Domain

	Accuracy		Response Time	
Variable	$\bar{X}$	SD	$\bar{X}$	SD
Electronics trouble-shooting: simple				
Trial 1	.77	.12	17.46	5.63
Trial 2	.85	.11	15.75	5.16
Trial 3	.89	.10	14.31	4.13
Trial 4	.91	.08	12.98	3.67
Trial 5	.93	.07	12.11	3.24
Trial 6	.93	.07	11.83	3.23
Trial 7	.94	.06	11.27	3.01
Trial 8	.95	.06	11.14	3.06
Average				
Electronics trouble-shooting: complex				
Trial 1	.75	.19	30.60	10.31
Trial 2	.80	.17	27.56	9.32
Trial 3	.82	.16	25.10	8.41
Trial 4	.84	.16	25.20	8.54
Trial 5	.83	.17	25.25	9.02
Trial 6	.83	.16	23.94	8.35

scoring degrees of rotation and number of surfaces carried, respectively, by two orthogonal polynomial contrasts each to represent the linear and the quadratic trends. For the position items, two orthogonal polynomials represented the linear and quadratic trends for degrees of rotation. The number of surfaces carried was not scored for position items, because no attaching process is postulated.

Distractor type was handled somewhat differently for response time data than for accuracy data. For accuracy data, the distractor type was directly included in a mathematical model for the prediction of individual subject responses to the 72 items by applying the linear logistic latent trait model (LLTM, Fischer, 1973). However, for response time data, distractor type was not included in the mathematical model. Individual response times are usually unreliable indicators of processing duration, so that averaging over tasks is necessary. Response times were averaged over distractor type because a preliminary analysis indicated that distractor type had no significant effect on response time.

Response time. Response times were modeled by a within-subjects analysis of variance with 582 subjects. The within-subjects analysis of variance can accommodate contrasts to represent the effects of the independent variables on the mean item response times and yet can also provide prediction at the individual subject level. An item response theory approach would have advantages over a within-subjects analysis of variance, but it could not be applied to the data in the current study. Although the mathematical modeling of response times in an item response theory model has some initial development (e.g., Scheibelchner, 1985), additional work on parameter estimation is needed for practical application.

The effects for the within-subjects analysis of variance included contrasts for time, to represent changes over occasions and the various contrasts described above. Surfaces and degrees are nested within problem type because degrees are scored separately by problem type and because surfaces do not apply to position items. For spatial items, contrasts for the surfaces-by-degrees interaction were added to the analysis to provide a full set of contrasts. Changes in the underlying cognitive processes were examined by comparing the interaction of the independent variables with time. Table 4 presents the degrees of freedom, *F* values, and probabilities associated with the effects in the within-subjects analyses of variance. It can be seen that both the time and the problem type main effects are significant. An inspection of the plots of response time by problem type and time, as well as the single degree-of-freedom tests underlying the main effects, indicated that response time decreases across occasions and that position items are solved more rapidly than spatial items. However, time interacts significantly with problem type. An inspection of the plots and the single degree-of-freedom contrasts revealed that the relative advantage of position items over spatial items increases over occasions.

Within spatial items, it can be seen that both the linear and quadratic effects of degrees and surfaces and the degrees-by-surfaces interaction were significant. An inspection of the plots of response time by degrees and surfaces

Table 4
Within Subjects Analysis of Variance
for Response Times at Three Occasions

Variables	df	F	p
Time	2,1162	630.03	<.000
Problem type	1,581	59.56	<.000
Time by problem type	2,1162	18.95	<.000
Spatial problems			
Degrees - linear	1,581	412.55	<.000
- quadratic	1,581	97.69	<.000
Surfaces - linear	1,581	532.09	<.000
- quadratic	1,581	176.30	<.000
Degrees by surfaces	4,2324	51.46	<.000
Time by degrees	4,2324	3.05	.016
Time by surfaces	4.2324	25.17	<.000
Time by surfaces by degrees	8,4648	7.62	<.000
Position problems			
Degrees - linear	1,581	31.64	<.000
Degrees - quadratic	1,581	130.65	<.000
Time by degrees	4.2324	7.92	<.000

indicates that response time increases monotically with both variables. However, the single degree-of-freedom tests underlying the degrees-by-surfaces interaction, and the plots, indicate that the effect of degrees of rotation generally is more linear for one-surface item than for two- or three-surface items.

The significant interactions of all independent variables with time for the spatial items indicate changes in construct representation across occasions. Both surfaces and degrees, and their interaction, interacted significantly with time. The nature of the changes across occasions was determined by inspecting the plots of response time by degrees and surfaces and by interpreting the single degree-of-freedom tests that underlie the significant interactions with time. In general, the effect of degrees of rotation on response time becomes more linear across occasions, due to a decrease in the quadratic trend. Further, degrees interact decreasingly with surfaces, such that the effect of degrees of rotation is more similar between one-surface items and two- or three-surface items. In contrast, the number of surfaces carried generally decreases in impact across occasions. Specifically, the one-surface items differ substantially less from the two- and three-surface items less on the last posttest than on the pretest or the first posttest.

Table 4 also shows the results for the position items. It can be seen that both the linear and quadratic trends of degrees are significant. An inspection of the plots revealed that, unlike spatial items, the effect was not monotonic because

90° had the shortest response time. Further, time has a significant interaction with the degrees effect, such that the quadratic trend increases.

Another type of analysis that is often applied to mathematically model response times is the regression of the item means on the independent variables. For this analysis, the 72-item means were regressed on the linear and quadratic trends for degrees (for both spatial and position items) and the linear and quadratic trends for surfaces, problem type, and distractor type within each occasion. The results indicated the same significant changes over occasions for the independent variables, as above. Further, the results indicated high predictability of response time means at all three occasions, as indicated by multiple correlations of .88, .88, and .84, respectively, for the pretest, first posttest, and second posttest.

Accuracy. The impact of the independent variables on the accuracy of spatial processing was determined by linear logistic modeling of item responses. The linear logistic latent trait model (LLTM; Fischer, 1973) is a constrained version of the Rasch model in which item difficulty is modeled from variables that reflect the impact of underlying components of the items. Thus, LLTM contains fewer parameters than the Rasch model, because item difficulties are replaced by values that are predicted from the underlying components. In the current study, the eight independent variables were the predictors of item difficulty in LLTM.

Table 5 presents the LLTM weights, η, and standard errors, σ_η from separate modeling of item difficulty within each occasion, linked across test order

Table 5
Linear Logistic Models of
Item Difficulty at Three Occasions

	Pretest		Posttest1		Posttest2		z	z
Variables	η	σ_η	η	σ_η	η	σ_η	Pre Post1	Post1 Post2
Distractor type	-.07^{+}	.04	-.15**	.04	-.17**	.05	1.43	.35
Problem type	-.10**	.05	-.48**	.05	-.64**	.05	5.35	2.25
Spatial problems								
Degrees								
Linear	.16**	.03	.14**	.03	.23**	.03	-.71	-2.14
Quadratic	.10**	.02	.10**	.02	.15**	.02	.00	-1.78
Surfaces								
Linear	.83**	.03	.72**	.03	.76**	.03	2.62	-.95
Quadratic	-.18**	.02	-.08**	.02	-.08**	.02	-3.57	.00
Position problems								
Degrees								
Linear	-.01	.05	.24**	.05	.14**	.05	-3.52	1.41
Quadratic	.19**	.03	.24**	.03	.15**	.03	-1.19	2.38
r_{Rasch}	.88		.88		.84			

Note. $^{+}$ p<.10. ** p<.01.

conditions. Program LINLOG (Whitely & Nieh, 1981) was used to obtain estimates for the parameters of LLTM. In LINLOG, indeterminancies of scale for both the Rasch model and any constrained LLTM model are resolved by setting the mean item difficulty to zero. For LLTM, the mean item difficulty that is reproduced by the model (i.e., the product of the complexity factor values and the factor weight) is set to zero. Error variances for the parameter estimates are obtained in LINLOG from the diagonal of the inverse of the information matrix, as in other applications of LLTM (see Fischer & Formann, 1982).

These LLTM model weights in Table 5 are analogous to regression weights and standard errors for modeling item difficulties, as scaled in the Rasch model. It can be seen that good fit was obtained at all occasions, because the correlations of the LLTM model item difficulties with the Rasch model item difficulties ranged from .83 to .88. The effects of the independent variables are similar to the response time models. The variables for processing on spatial items, degrees of rotation, and number of surfaces carried were significant for both the linear and the quadratic trends at all occasions. The trend variables for attaching on position problems, degrees of rotation, also were significant at all occasions, although the linear effect was not significant on the pretest.

Unlike the response time models, the contrast for type of distractors was clearly significant at the posttests and marginally significant at the pretest. Distractor sets with matched orientations to the key were processed less accurately than distractor sets with mixed orientations. Similarly, the contrasts for position versus spatial problems were significant at all occasions.

Under certain circumstances, it is possible to test the similarity of LLTM models across occasions by coding product interaction vectors for each independent variable (e.g., see Fischer & Formann, 1982). However, this method requires modeling items across occasions with a single ability. Because dimensionality changes across occasions in the current data as postulated in MRMLC and because 72 items on a single test are numerically cumbersome for CML estimation, the unified analysis was not attempted. However, z tests for pairwise comparisons of LLTM estimates are reported, using the pooled standard errors, in Table 5. Using a significance level of .05, with a critical two-tailed z of 1.96, it can be seen that distractor type does not vary significantly in effects across occasions. However, the position versus spatial contrast increases significantly in impact over occasions. For spatial problems, the surfaces effect decreases significantly from the pretest to the first posttest but not between the two posttests. For degrees of rotation, the linear component increases significantly from the first to the second posttest. For position items, the linear trend for degrees of rotation increases significantly from the pretest to the first posttest while the quadratic effect increases between the two posttests.

An inspection of the plots of item difficulty by the independent variables reveals that on the pretest one-surface items are much less difficult than two- or three-surface items, which differ little from each other. On the posttests, the difference in difficulty between one- and two- or three-surface items is smaller

while the difference between two- and three-surface items increases slightly. Separate plots for spatial and position items reveal that 90° rotations are much easier for position items at all occasions. The effect is less pronounced for spatial items. Further, the change in item difficulty from pretest to posttest is greater for position items than for spatial items, and this fact reflects the increased impact of the contrast for position versus spatial items.

Nomothetic span. The nomothetic span of initial spatial ability and the two modifiabilities were examined by their predictions of both overall learning and the course of learning in the applied knowledge domain. Two indexes of learning, accuracy, and response time per correct decision were examined for both the simple and the complex learning task.

The significant multiple correlations in Table 6 indicate that both dependent variables within both the simple and the complex learning task were predicted significantly by initial spatial ability and the two modifiabilities. An inspection of the standardized regression coefficients reveals that initial ability had a significant independent contribution in predicting all dependent variables, as expected. However, the significant standardized regression coefficients for the first modifiability also had a significant independent contribution in predicting all dependent variables. Thus, incremental validity is supported for the first modifiability. The second modifiability did not reach significance in any regression equation.

The role of spatial ability and of modifiability in predicting the process of learning was examined by structural equation modeling of the covariances between trial accuracies and the abilities. An empirically plausible model of individual differences in learning is needed prior to determining the relationship of initial ability and modifiability to learning. Thus, individual differences in learning were modeled prior to modeling the relationship of ability and modifiability to learning.

For both the simple and complex learning tasks, the covariances among trials were modeled by a Wiener simplex model (see Jöreskog, 1970). Because the variances are decreasing for both the simple and complex learning task (see

Table 6
Prediction of Learning from a Dynamic Spatial Test

	Standardized Regression Coefficients				
Variables	Initial Ability	Modif 1	Modif 2	R	F
Simple learning task					
Accuracy	.29**	.20**	.01	.35**	13.92
Response time	-.16**	-.14*	-.11+	.21**	4.41
Complex learning task					
Accuracy	.36**	.22*	.03	.43**	11.90
Response time	-.17*	-.20**	-.06	.26**	3.72

Note. + p<.10. * p<.05. ** p<.01.

Table 2), a reverse Wiener simplex was used to model the trial covariances. Thus, the factorial complexity of performance decreases over trials, presumably due to mastery of factors that contributed to individual differences (cf., Jones, 1970). That is, if for person j on trial t x_{jt} is the performance on trial t, f_{jt} is the score on the factor for trial t and e_{jt} is the error on trial t. The following equation expresses the reverse Wiener simplex model for five trials:

$$\begin{aligned} x_1 &= f_1 + f_2 + f_3 + f_4 + f_5 + e_1 \\ x_2 &= f_1 + f_2 + f_3 + f_4 + e_2 \\ x_3 &= f_1 + f_2 + f_3 + e_3 \\ x_4 &= f_1 + f_2 + e_4 \\ x_T &= f_1 + e_5 \end{aligned} \tag{2}$$

In this model, f_1 is a general learning factor because it is involved in all trials. In contrast, f_5 is an early learning factor because it drops out on Trial 2 while f_2 is a late learning factor that persists until Trial 4.

Jöreskog's structural equation model for the Wiener simplex has the following form:

$$\Sigma = \Lambda\Phi\Lambda' + \Psi, \tag{3}$$

where Σ is the $T \times T$ covariance matrix between T trials, Λ is the $T \times T$ factor loading matrix, Φ is the diagonal $T \times T$ factor covariance matrix, and Ψ is the diagonal $T \times T$ matrix of error variances. The factor loading matrix, Λ, contains only fixed values of 0 or 1 to define the reverse Wiener simplex pattern, as follows:

$$\Lambda = \begin{matrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{matrix}. \tag{4}$$

For both tasks, the trial covariances were modeled by successively complex reverse Wiener simplex models, which were implemented in Program EQS for structural equation models. In the least complex simplex model, M1, Λ was constrained as in (4), Φ is a diagonal matrix with no factor covariances, and Ψ is a diagonal matrix with all error variances constrained to a common value. M1 is an extremely parsimonious account of trial covariances, as the common error constraint makes the model even less complex than (2), which has separate errors for each trial. After determining an adequate model for the trial covariances, the covariances for the full set of variables, including abilities as well as trials, were modeled.

Table 7 shows the goodness-of-fit test, significance level, and Bentler-Bonnett nonnormed fit index for models of the simple learning task. A preliminary model, independence, was highly significant ($p < .001$), which indicates significant covariances between the eight trials. The first structural

Table 7
Structural Equation Modeling of Simple Learning Task

Model	df	χ^2	p	Bentler-Bonnett Normed Fit
Trial models				
Independence	28	1666.20	<.001	--
Simplex models				
M1-constrained errors	27	71.50	.001	.972
M2-partial constrained errors	25	40.76	.022	.989
M3-1 factor covariance	24	30.52	.168	.995
M4-2 factor covariances	23	19.65	.662	1.002
Trial and ability models				
M5-null model	47	83.24	<.001	.971
M6-general learning predicted	44	60.32	.051	.998
M7-general learning and three stages predicted	41	36.00	.692	1.004

model for the trials, M1, did not fit the data. In M2, constraints on error variances were released only if indicated by the LaGrange Multiplier (LM) test as improving fit significantly. Although fit improved over M1 by releasing two constraints on error variances, M2 still did not fit, so, in M3 and M4, factor covariances were estimated as indicated by the LM test. M4, with two factor covariances, did fit the data. It should be noted that M4 still provides a highly parsimonious account of the trial covariances, as it contains only four more parameters than M1.

Table 7 also shows the series of models in which spatial ability and the two modifiabilities were added to the trial covariance matrix. The null model (M5), in which abilities were not permitted to correlate with learning, did not fit the data. In M6, only the general learning factor (f_1) is regressed on the abilities. M6 fit the data. However, the LM test indicated that allowing further regression of the learning factors on abilities would improve fit. This is shown as M7, which fits the data. The Bentler-Bonett fit index is at the upper boundary. The LM test indicated no further improvements in fit from releasing permissible constraints.

Figure 2 shows the structural equations between the abilities and the learning factors. It can be seen that initial ability had significant weights for predicting general learning and middle-stage learning (f_4). An inspection of the parameter estimates indicated that, although the general learning factor has a large variance, ($s^2_{f1} = 27.14$) the middle-stage learning factor has a relatively small variance ($s^2_{f4} = 3.54$). The first modifiability had significant weights in predicting both general learning and an empirically important early learning

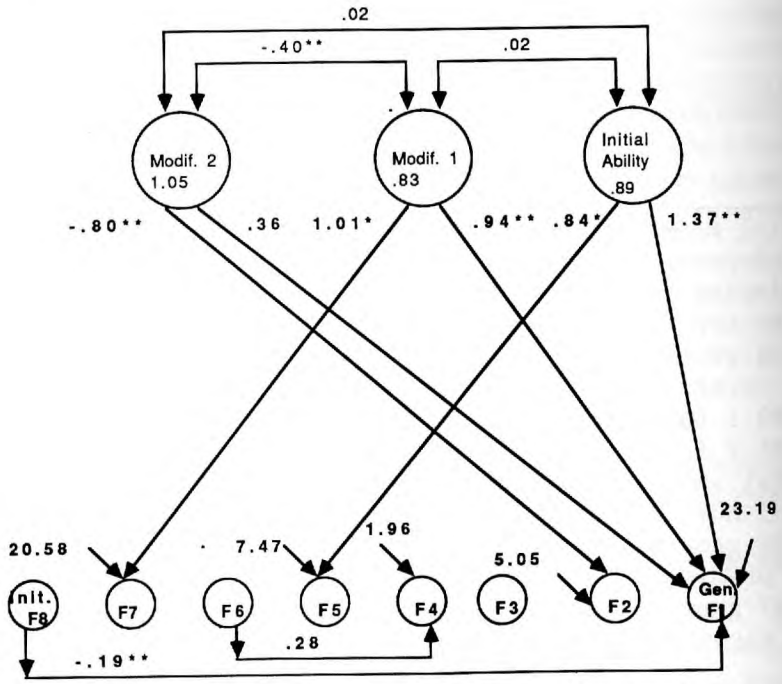

FIGURE 2. *Final structural equation model for simple learning task*

factor, f_7, which has a relatively large variance ($s^2_{f7} = 22.24$). The second modifiability had a significant negative weight in predicting late learning, which has a relatively small variance ($s^2_{f6} = 6.09$). Figure 2 also shows that the two significant trial correlations are small ($r_{f8,f1} = -.19$ and $r_{f4,f6} = .28$). Thus, the departure from the model with no covariances is minimal.

Table 8 presents results for similar structural equation models of the complex learning task. For the trial covariances, the preliminary model (independence) indicates that the covariances between the six trials are highly significant. The first reverse Wiener simplex model, with six learning factors, failed numerically, and one factor variance had to be constrained to zero to empirically identify the model. Thus, the first structural equation model, M1, is the reverse Wiener simplex with five learning factors, with constrained error variances. In M2, two error constraints are released (as indicated by the LM test), but the model still did not fit, and one factor variance was not significant. Thus, M3 and M4 contain only four learning factors. M3, with no factor covariances, did not fit, but releasing one covariance between the late and initial learning factors in M4 led to a model that fit.

Figure 3 shows M4, which is a highly parsimonious account of the data because there are few free parameters. It can be seen that the general learning factor has a substantially larger variance than the other factors. Late learning, in contrast, has the smallest variance. However, the last trial has a substantially

Table 8
Structural Equation Modeling of Complex Learning Task

Model	df	χ^2	p	Bentler-Bonnett Normed Fit
Trial models				
Independence	15	1127.60	<.001	--
Simplex models				
M1-constrained errors	15	34.82	.002	.982
M2-partial constrained errors	13	26.59	.014	.986
M3-1 reduced learning factors	14	26.83	.020	.988
M4-1 factor covariance	13	21.31	.067	.993
Trial and ability models				
M5-null model	31	74.28	<.001	.957
M6-general learning predicted	28	44.12	.027	.982
M7-general learning and three stages predicted	27	34.99	.138	.991

larger unique (remainder) variance, which indicates that substantial individual differences remain.

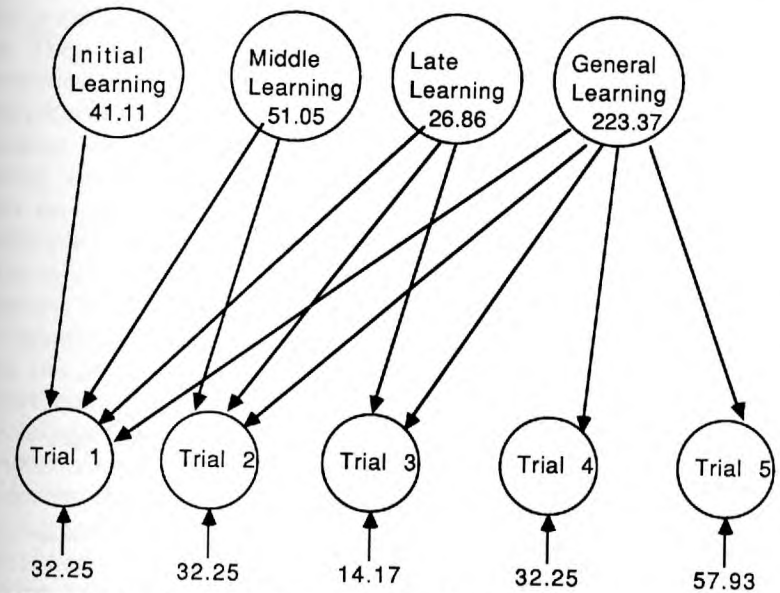

FIGURE 3. *Structural equation model of trials only for complex learning task*

Table 8 also presents the structural equation models with the ability variables added to the best trial model, M4. The null model (M5), in which the learning factors are not regressed on ability, did not fit the data. Allowing the first learning factor to be regressed on the three abilities significantly improved fit in M6 ($\Delta\chi^2 = 30.16$, $df = 3$, $p < .001$). However, the model did not fit, and the LM test indicated that regressing the remainder factor on the last trial on the second modifiability would improve fit. This model, M7, fit the data, and the LM test indicated no further significant regression of learning on ability. Figure 4 presents the model for M7.

Discussion

The goal of this study was to examine a cognitive modifiability measure by applying some new methods to remedy problems that have limited the potential of other validity studies. In the current study, spatial ability and two modifiabilities were estimated within an item response theory model specifically formulated for learning and change (Embretson, 1991a). The estimates were derived from repeated measurements of ability within a single 45-minute testing session. The modifiabilities were associated with two different interventions, cued training with a physical analogue of the mental task, and strategy training on the cognitive model underlying the task. The descriptive statistics indicated that the interventions were associated with substantial increases in performance on the spatial ability task with the total standardized effect estimated to be .83. The standardized effect for the first modifiability alone was .55.

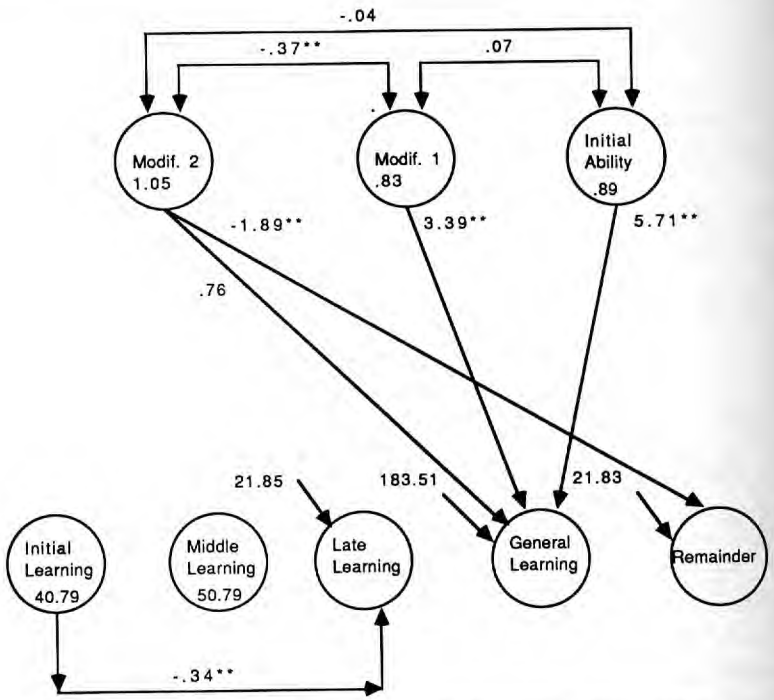

FIGURE 4. *Final structural equation model for complex learning task*

These substantial effect sizes are comparable to other studies on dynamic testing. For example, Budoff and Hamilton's (1974) effect size ranged from .46 to 1.20 while Carlson and Weidl's effect size ranged from .60 to 1.77. Embretson's (1987) effect size of .64 for cued training intervention for spatial ability is quite similar to the results obtained here.

The most important issue in the current study was to examine the criterion-related validity of cognitive modifiability. In the current study, learning was precisely measured in several stages of mastering an applied knowledge domain. The applied knowledge domain was a technical skill for which the examinees had little or no prior knowledge. The mean accuracy and response time on the learning criterion were significantly predicted by initial spatial ability and the two modifiabilities. Although the multiple correlations of learning with the abilities were modest, they are comparable in magnitude to other validity results that report attenuated correlations (cf., Anastasi, 1987; Hunter, Schmidt, & Hunter, 1979).

Most importantly, however, the incremental validity for the first modifiability, following the physical analogue training, was supported for all dependent variables. That is, the first modifiability had a significant regression weight, along with initial ability, for predicting both accuracy and response time for both the simple and complex learning task. Although the standardized regression coefficients were not as high as for initial ability, the effects were still substantial.

In contrast, incremental validity for the second modifiability, following strategy training, was not supported for predicting overall learning on either task. These results on the second modifiability should not necessarily be interpreted as the failure of cognitive modifiability based on strategy training to increase criterion-related validity, however. The potential of strategy training may have been limited by its order in the testing sequence. That is, strategy training followed analogue training, which already had achieved substantial effect on performance levels. Or, the incremental validity of the second modifiability may be evident only in certain stages of learning.

The validity of the spatial abilities to predict the process of learning was also examined. Prior to investigating validity, the nature of individual differences in the learning process was determined. For both the simple and complex learning tasks, the process of individual differences in learning was fit by a reverse Wiener model. In the reverse Wiener simplex model, performance has become decreasingly complex factorially as successively more elements that contribute to individual differences have been mastered. The modified versions of the Wiener model allowed some covariances between learning factors. In the simple learning criterion task, only two covariances between the learning factors were needed to achieve model fit, and, in the complex learning criterion, only one factor covariance was required. Because error variances were also highly constrained, the final models should be regarded as a highly parsimonious account of the learning trial covariances.

The most clear results on predicting the process of learning were obtained for the general learning factor. For both the simple and complex learning task, initial spatial ability and the first modifiability had significant regression coefficients. Thus, because the general learning factor remained important at all stages according to Wiener simplex model, initial ability and the first modifiability predicted the process of learning.

The results also provided some support for initial ability and the modifiabilities' predicting individual differences in learning at some specific stages as well. For the simple learning task, initial ability and the second modifiability additionally predicted relatively small middle-stage and late-stage learning factors, respectively, while the first modifiability additionally predicted a large early learning factor. Thus, the first modifiability had a relatively important role in predicting early learning. For the complex learning task, the general learning factor was predicted by ability and the modifiabilities, but no other learning factors were predicted significantly. However, the second modifiability had a significant negative regression coefficient for predicting remaining individual differences on the last trial. The remainder factor was relatively large, and partially consisted of errors, but it had to consist of unlearned factors as well.

The process of learning should be interpreted with respect to the type of learning criterion task used in the current study. That is, performance in acquiring the applied knowledge domain, even at the first stage, probably maximally involves an individual's skills and strategies for acquiring new knowledge, because the role of prior knowledge is minimal. Thus, modifiability could be expected to be involved in the general learning factor and in early learning, as was found for the first modifiability. However, if the role of prior knowledge had been greater initially, perhaps modifiability would emerge as a predictor only at specific stages of learning.

To explain the various support for the incremental validity of modifiability that was obtained, it is necessary to consider how the intervention influenced construct representation in terms of the strategies, processes, and knowledge structures that were involved in performance. As expected, the results from mathematically modeling both response time and accuracy supported the empirical plausibility of the cognitive model that was postulated for the spatial task. Good fit was achieved for the mathematical models, and the independent variables that were postulated to influence attaching and folding were significant. Further, the results clearly indicated that construct representation changes from the pretest to the two posttests. More specifically, the results indicated changes across occasions in (a) the strategies applied to spatial versus position items, (b) the attaching process for both spatial and position items, and (c) the folding process for spatial items. These will be discussed in turn.

First, several results indicated that strategy differences increased between spatial and position items across occasions. General strategy differences between spatial and position items were supported by finding that position items were solved both more rapidly and more accurately than spatial items. Further, the trend for degrees of rotation on response time differed between position items and spatial items. These results were consistent with Embretson and

Waxman's (1989) model in which position items were hypothesized to not require a folding process. The increasing differences in both response time and accuracy between spatial and position problems across occasions suggested increasingly differentiated strategies for applying the folding process.

Second, the attaching process became more stable across occasions. For both spatial and position problems, degrees of rotation had increasing impact on both response time and accuracy. However, the nature of the effects differed between spatial and position problem, and between response time and accuracy. For spatial problems, degrees of rotation had a more linear effect on response time on the posttests and had a more similar effect over items with a different number of surfaces to be carried. For position problems, degrees of rotation had an increasingly quadratic effect on response time. For both problem types, degrees of rotation had a quadratic effect on accuracy.

Third, the folding process also changed across occasions. The quadratic trend decreased for both accuracy and response time, which was due to a lessening of the difference between one-surface items versus two- or three-surface items. A plausible explanation for these findings is that there are qualitative differences in applying the folding process across occasions. One-surface items, in contrast to two- or three-surface items, can be solved by pairwise comparisons of the markings on the sides, because the three sides on the folded correct answer are also adjacent on the stem. In contrast, two- or three-surface items cannot be solved by pairwise comparisons. The larger difference between one-surface versus two- or three-surface items on the pretest (i.e., the quadratic effect of surfaces) indicates initial inconsistencies in applying the folding process. On the posttests, the difference between one-surface problems and the two- and three-surface problems decreased, which is consistent with a greater continuity of processing. Similarly, the increased similarity of the effect for degrees of rotation on response time between one-surface versus two- or three-surface items also indicates greater continuity of processing.

In conclusion, the stabilization of processing and the development of differential strategies are central to understanding modifiability in the current study. In the spatial folding task, because folding is postulated to follow attaching, folding involves further mental transformations of the stem while preserving the results from the attaching process. The requirement of preservation of preceding transformations under further transforms has often been hypothesized as central to spatial visualization ability (e.g., Just & Carpenter, 1985; Mumaw & Pellegrino, 1984; Poltrock & Brown, 1984), and so it is not surprising that it emerges here as a major factor. However, what is unique to the current study is that measuring individual differences in modifiability, which reflects in part the increased stabilization and differentiation of the folding process in the posttests, leads to greater criterion-related validity.

The results generally indicate, then, that greater validity can be obtained when processing strategies for spatial visualization ability are more stable and differentiated. One implication that could be drawn is that a better measure of ability can be obtained by providing more practice prior to measurement.

Because individual differences on the posttests in the current study were influenced by the sum of initial ability and modifiability, posttest scores will tap both and thus provide higher criterion-related validity. Thus, perhaps supplying a few more practice items prior to measurement would lead to increased validity.

However, simply providing more practice items may not provide more valid measurement for several reasons. First, the increased validity in the current study was derived from posttests that were preceded by very extensive practice. Even the first posttest in the current study is preceded by 28 pretest items, 15 intervention items, and several examples in the initial instructions. It is not clear that increased validity could be obtained from any less extensive practice, and that, if not, using the pretest items to provide another ability estimate costs nothing in subject time. Second, retaining both initial ability and modifiability as separate measurements could lead to the highest validity, because they may be differentially weighted in predicting different learning criteria. In fact, some support for differential prediction at different stages of learning is found in the current study, but prediction also could vary with the type of learning criterion. That is, for predicting some criteria, initial ability may have the highest weight while for predicting other criteria modifiability may have the highest weight. Third, it is not clear that simple practice will provide the changed construct representation or nomothetic span that was observed in the current study. In a previous study, Embretson (1987) found that simple practice led to about half the increase in performance that was observed for the physical analogue training. It is possible that similar changes in construct representation will be observed with simple practice. However, the effects may require too much time to be practically feasible.

In summary, the results supported the construct and criterion-related validity for the cognitive modifiability of spatial visualization items. The application of some relatively new methods for estimating and validating modifiability led to positive support for modifiability as a direct measurement of learning ability.

References

Anastasi, A. (1987). *Psychological testing.* New York: Macmillan.

Babad, E. Y., & Budoff, M. (1974). Sensitivity and validity of learning-potential measurement in three levels of ability. *Journal of Educational Psychology, 66,* 434–447.

Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 3–20). Madison, WI: University of Wisconsin Press.

Bransford, J. P., Delclos, V. R., Vye, N. J., Burns, M. S., & Hasselbring, T. S. (1987). State of the art and future directions. In C. S. Lidz (Ed.), *Dynamic assessment* (pp. 479–496). New York: Guilford.

Budoff, M. (1987). The validity of learning potential assessment. In C. S. Lidz (Ed.), *Dynamic assessment* (pp. 52–81). New York: Guilford.

Budoff, M., & Hamilton, J. L. (1974). Learning potential among the moderately and severely mentally retarded. *Mental Retardation, 12,* 33–36.

Campione, J. C., & Brown, A. (1987). Linking dynamic assessment with school achievement. In C. S. Lidz (Ed.), *Dynamic assessment* (pp. 82–115). New York: Guilford.

Carlson, J. S., & Weidl, K. H. (1979). Toward a differential testing approach: Testing-the-limits employing the Raven's matrices. *Intelligence, 3,* 323–344.

Cronbach, L. J., & Furby, L. (1970). How should we measure change—or should we? *Psychological Bulletin, 74,* 68–80.

Dearborne, D. F. (1921). Intelligence and its measurement: A symposium. *Journal of Educational Psychology, 12,* 123–147, 195–216.

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

Embretson, S. E. (1987). Improving the measurement of spatial aptitude by a dynamic testing procedure. *Intelligence, 11,* 333–358.

Embretson, S. E. (1989). *Measuring learning ability by dynamic testing* (Final Report for Air Force Contract No. AFOSR-88-0242). Lawrence, KS: University of Kansas, Department of Psychology.

Embretson, S. E. (1991a). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56,* 495–515.

Embretson, S. E. (1991b). Implications of a multidimensional latent trait model for measuring change. In L. Collins & J. Horn (Eds.), *Best methods of analyzing change* (pp. 184–197). Washington, DC: American Psychological Association Books.

Embretson, S. E., & Waxman, M. (1989, November). *Models for processing and individual differences in spatial folding.* Paper presented at the Annual Meeting of the Psychonomic Society, Atlanta.

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

Fischer, G. H., & Formann, A. K. (1982). Some applications of logistic latent trait models with linear contraints on the parameters. *Applied Psychological Measurement, 6,* 397–416.

Glaser, R., & Bassok, M. (1989). Learning theory and the study of instruction. In M. R. Rosenzweig & L. W. Porter (Eds.), *Annual review of psychology* (pp. 631–666). Palo Alto, CA: Annual Reviews.

Guthke, J. (1982). The learning test concept—An alternative to traditional static tests. *The German Journal of Psychology, 6,* 306–324.

Hunter, J. E., Schmidt, F., & Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin, 86,* 721–735.

Jones, M. (1970). A two-process theory of individual differences in motor learning. *Psychological Review, 77,* 353–360.

Jöreskog, K. G. (1970). Estimation and testing of simplex models. *British Journal of Mathematical and Statistical Psychology, 23,* 121–145.

Just, M., & Carpenter, P. (1985). Cognitive coordinate systems: Accounts of mental rotation and individual differences in spatial ability. *Psychological Review, 92,* 137–172.

Kyllonen, P. C., & Christal, R. E. (in press). In R. Dillon & J. W. Pellegrino (Eds.), *Testing: Theoretical and applied issues.* San Francisco: Freeman.

Kyllonen, P. C., & Shute, V. J. (in press). Learning indicators from a taxonomy of learning skills. In P. L. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), *Learning and individual differences.* New York: Freeman.

Kyllonen, P. C., & Stevens, D. (in press). Cognitive abilities as determinants of success in learning to trouble-shoot. *Learning and Individual Differences.*

Lidz, C. (1987). *Dynamic testing.* Beverly Hills, CA: Guilford.

Mumaw, R. J., & Pellegrino, J. W. (1984). Individual differences in complex spatial processing. *Journal of Educational Psychology, 76,* 920–939.

Poltrock, S. E., & Brown, P. (1984). Individual differences in visual imagery and spatial ability. *Intelligence, 8,* 93–138.

Rand, Y., & Kaniel, S. (1987). Group administrating of the LPAD. In C. S. Lidz (Ed.), *Dynamic assessment* (pp. 196–214). New York: Guilford.

Scheibelchner, H. (1985). Psychometric models for speed-test construction: The linear exponential model. In S. E. Embretson (Ed.), *Test design development in psychology and psychometrics* (219–244). New York: Academic.

Stake, R. E. (1961). Learning parameters, aptitudes and achievements. *Psychometric Monographs* (Whole No. 9).

Tzuriel, D., & Klein, P. S. (1987). Assessing the young child: Children's analogical thinking modifiability. In C. S. Lidz (Ed.), *Dynamic assessment* (pp. 268–287). New York: Guilford.

Vale, D. (1984). *MICROCAT testing system.* St. Paul, MN: Assessment Systems Corp.

Vye, N. S., Burns, M. S., Delclos, V. R., & Bransford, J. D. (1987). A comprehensive approach to assessing intellectually handicapped children. In C. S. Lidz (Ed.), *Dynamic assessment* (pp. 327–359). New York: Guilford.

Whitely, S. E., & Nich, K. (1981). *Program MULTICOMP.* Unpublished manuscript.

Woodrow, H. (1938). The relationship between abilities and improvement with practice. *Journal of Educational Psychology, 29,* 215–230.

Woodrow, H. (1946). The ability to learn. *Psychological Review, 53,* 147–158.

Author

SUSAN E. EMBRETSON is Professor, Department of Psychology, University of Kansas, Lawrence, KS 66045. *Degree:* PhD, University of Minnesota. *Specializations:* psychometric methods, multivariate statistics, and intelligence and cognition.