Embretson
Large-scale Assessments in Education (2023) 11:9
https://doi.org/10.1186/s40536-023-00158-8

Large-scale Assessments in Education

RESEARCH | Open Access

Understanding examinees' item responses through cognitive modeling of response accuracy and response times

Susan Embretson*

*Correspondence: susan.embretson@psych.gatech.edu
Georgia Institute of Technology, Atlanta, Georgia, USA

Abstract

Understanding the cognitive processes, skills and strategies that examinees use in testing is important for construct validity and score interpretability. Although response processes evidence has long been included as an important aspect of validity (i.e., *Standards for Educational and Psychological Tests*, 1999), relevant studies are often lacking, especially in large scale educational and psychological testing. An important method for studying response processes involves explanatory mathematical modeling of item responses and item response times from variables that represent sources of cognitive complexity. For many item types, examinees may differ in strategies applied to responding to items. Mixture class item response theory models can identify latent classes of examinees with different processes, skills and strategies based on their pattern of item responses. This study will illustrate the use of response times in conjunction with explanatory item response theory models and mixture models, to provide information relevant to test validity and, hence, to score interpretations.

Introduction

Understanding the response processes of examinees has had increased interest for high-stakes educational and psychological testing due to the increasing availability of item response time and log file data for examinees. Studying response processes in testing not only provides important data for test validity and score interpretations, but also can lead to important changes in test content and item design. At issue, however, is applying appropriate methods to achieve these goals in the context of standard testing procedures and scoring.

Mathematical modeling of response accuracy and response time is a method used in a large percentage of cognitive studies to understand response processes and their interaction in task performance (Busemeyer & Diederich, 2010). The models typically are applied in the context of theory-based manipulations of the tasks.

Although not routinely applied, models for understanding cognitive processes have long been available for educational and psychological testing. In the context of item response theory (IRT), several explanatory models using theory-based predictors, have been developed. Fischer (1973) developed the linear logistic test model (LLTM, 1973) for binary response data to predict item difficulty from variables representing sources of processing difficulty. The linear logistic partial credit model (LPCM; Fischer & Parzer, 1991) was later developed to accommodate polytomous response items. As summarized later in this article, many other IRT-based models have been developed in the last several decades to examine response processes. This includes models to jointly predict response accuracy and response times from theory-based variables (e.g., Klein Entink et al., 2009) as well as models to examinee strategy differences (e.g., von Davier & Rost, 1995).

Janssen (2016) presented applications of LLTM and related models to a wide variety of tests, including mathematics, reasoning, reading, science, personality and emotions. However, noticeably absent was routine applications for large scale testing. But, that limitation could change with the current interest in understanding response processes in large scale testing.

In this paper, a study on IRT modeling of cognitive processing for an aptitude test will be presented to illustrate the potential of understanding responses processes. However, some background will be presented first to place the study in context. First, an integrated version of the validity concept will be presented to show how studying response processes can impact the various aspects of validity as well as impact item and test design. The results from the study to be presented are directly relevant to three of the five aspects of validity. Second, a brief review of IRT models used in the current study to understand response processes will be presented.

Background

Validity is a major component of *the Standards for Educational and Psychological Tests*, including the most recent version (2014). However, the validity concept has changed substantially over time. Particularly important is the inclusion of response processes as a major component of validity. However, these changes are not well understood as many research articles and some textbooks still include an earlier version of the standards. Further, even in the most recent revision of the test standards, the methods listed for studying response processes are not sufficient. In this section, the earlier validity concepts first will be briefly described. Then, the current formulation will be presented, along with a consideration of methodologies for studying response processes. Finally, an integrated model of validity will be presented to show the impact of research on response processes on the various aspects of validity as well as on item and test design.

Validity and response processes

Validity has been a major aspect of testing standards since the *American Psychological Association Committee* on Psychological Tests met in 1954. Subsequent versions of the standards involved three organizations; *American Psychological Association*, *National Council on Measurement in Education* and *American Educational Research Association.* The first standards included four *types* of validity: (1) content validity, the representativeness of specified content areas in a test, (2) concurrent validity, the correlation of the test with examinees' current standing on some external variables, (3) predictive validity, the prediction of examinees' standing on some external variables and (4) construct validity, the degree to which test scores can be interpreted as reflecting the latent trait(s) presumed to underlie test performance.

The validity concept was relatively unchanged until 1999. That is, in 1999, the validity concept in the *Standards for Educational and Psychological Tests* changed substantially from other versions, which continues in the most recent version of the standards (2014). Unfortunately, many researchers, academics and textbooks still refer to the four separate types of validity.

In the current *Standards*, validity is a unitary concept where "the term construct is used in the *Standards* to refer to the concept or characteristics that a test is designed to measure" (*Standards*, 2014, p. 11). However, there are five sources of evidence for a construct: (1) Content, the knowledge, skills and attributes represented on the test, (2) Response Processes, the cognitive processes engaged in by examinees when responding to items, (3) Internal Structure, the interrelationships between items and test dimensionality, (4) Relationship to Other Variables, such as other constructs, criteria and examinee background variables and (5) Consequences, impact of test use on examinees from varying backgrounds. According to the standards, evidence for the various aspects should be appropriate for the construct that the test is designed to measure.

It has been sometimes argued that only one aspect of validity may be relevant to a particular test. In the context of achievement tests, Lissitz and Samuelson (2007) argue that the content aspect overshadows the other aspects for relevancy in score interpretations. In response to that view, an integrated model of validity was developed (Embretson, 2007, 2017). Figure 1 shows a somewhat revised version of the model to highlight the impact of empirical research on response processes. The five aspects of validity are organized as involving internal versus external aspects of a test. These represent test score meaning and significance, respectively. Further, the aspects of validity are interrelated in a causal manner, starting with the Content aspect. That is, the Content aspect concerns the representation on the test of the skills, attributes and knowledge through the features of items. In achievement testing, items are judged by experts for representing not only blueprint categories that include content areas and complexity levels, but also as involving appropriate cognitive complexity for

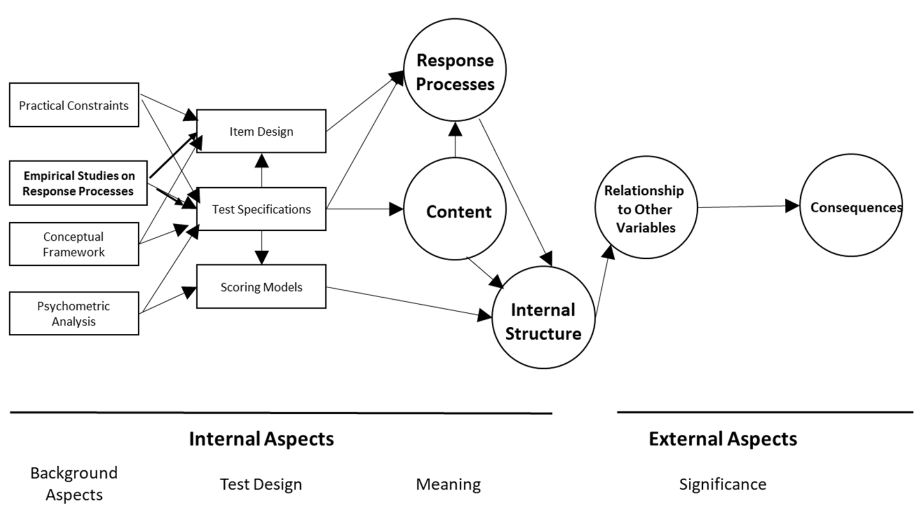

Fig. 1 An integrated model of construct validity

solution. The Content aspect also includes test administration and scoring conditions, as indicated in the *Standards* (2014).

The Response Processes aspect, as noted above, directly concerns the cognitive activities of examinees in responding to the items. As in cognitive psychology, cognitive activities in test items are directly driven by their features. Hence, the Content aspect has a direct relationship to the Response Processes aspect. Further, both the Content aspect and the Response Processes aspect drive the Internal Structure and the Relationship to Other Variables aspects of validity. That is, the representation of various content and cognitive activities by item features determines item intercorrelations and dimensionality, as well as test score relationships with other tests, criteria and background variables. Finally, those relationships impact the Consequences aspect.

The five aspects of validity evidence, in turn, are driven by three aspects of Test Design; Item Design, Test Specifications, and Scoring Models. Test Specifications include both the features of items (blueprints, item types, etc.) and the conditions of test administration (i.e., instructions, test presentation mode, etc.), as well as the type of scores to be extracted from items. Thus, Test Specifications determine both Item Design and Scoring Models.

Finally, Studies on Response Processes are conceptualized as part of the background variables that are related to the Test Design variables. Importantly for the current article, Studies on Response Processes can have impact on all three aspect of test design to assure that the observed processes are consistent with the intended interpretations of the test. That is, construct-relevant processes, skills and knowledge are being measured. Studies on Response Processes also can impact the prediction of item parameters. In fact, parameter predictability from item features that impact cognitive processing can provide a foundation for automatic item generation.

Psychometric models for response processes

As mentioned above, mathematical modeling as a major method for studying response processes in cognitive psychology. Busemeyer and Diederich (2010) present a logistic model for estimating dynamic signal detection as an example:

$$P(R_s = 1) = \frac{1 + e^{-2d(\theta+\beta)}}{1 + e^{-4d(\theta)}}$$

where R_s = response of signal detected, d = discrimination (evidence), β = response bias and θ = boundaries of speed-accuracy tradeoff. The logistic model format and symbols are similar to item response theory (IRT) models although, of course, the meaning is different.

The linear logistic test model (LLTM; Fischer, 1973) was available very early in the development of IRT models. LLTM is a Rasch-family IRT model for item responses that can be used to estimate the impact of various sources of cognitive complexity on item difficulty. That is, item difficulty is predicted from scores based on features of the items. In LLTM, the probability that person *j* solves item *i*. $P(X_{ij} = 1)$, is given as follows:

$$P\left(X_{ij} = 1\right) = \frac{\exp\left(\theta_j - \sum_k \tau_k q_{ik} + \tau_0\right)}{1 + \exp\left(\theta_j - \sum_k \tau_k q_{ik} + \tau_0\right)}, \quad (1)$$

where q_{ik} = score for item *i* on predictor *k*, τ_k = weight for predictor *k*, and τ_0 = intercept. Thus, predicted item difficulty, β_i', is given as follows:

$$\beta'_i = \sum_k \tau_k q_{ik} + \tau_0. \quad (2)$$

For example, for mathematical achievement items the values of q_{ik} can represent scores for the item on sources of computational, analytic and verbal (i.e., as in word problems) complexity that impact cognitive processing. LLTM is applicable to binary item response accuracy data. However, polytomous data, such as item scores, can be accommodated in the same manner with the linear partial credit model (LPCM; Fischer & Parzer, 1991).

A wide variety of models that are useful for studying cognitive processing using response accuracy data were developed subsequently. An overview of several models that can reflect cognitive processes based on response accuracy, denoted as explanatory IRT models, was presented by De Boeck and Wilson (2004, 2016). Several models can reflect individual differences in response patterns during testing. These include multicomponent models (e.g., Embretson, 1983; Embretson & Yang, 2013) to measure two or more traits based on processing differences between items, dynamic interaction models (e.g., Meulders & Xie, 2004) to reflect changes during testing, a random weights LLTM (Rijmen & De Boeck, 2002) to reflect person differences in the predictor weights for item difficulty, and a mixture distribution model (Rost & von Davier, 1995) to identify individual differences in item response strategies.

For example, Rost and von Davier's (1995) mixture distribution model identifies latent classes of examinees to reflect varying response strategies. That is, the separate latent classes, with different patterns of item difficulty, are developed to increase person fit. The mixture IRT model is given as follows:

$$P(\theta) = \sum_g \pi_g \frac{\exp\left(\theta_{jg} - \beta_{ig}\right)}{1 + \exp\left(\theta_{jg} - \beta_{ig}\right)}, \quad (3)$$

where θ_{jg} = trait level of person *j* in group *g*, β_{ig} = difficulty of item *i* in group *g* and π_g = probability for group *g*. Notice that the group subscript for item difficulty leads to different patterns of values within the groups.

Models for response time data have also been developed. Response time data is often highly skewed, so the models often involve log response time. Response time data can be modeled independently from response accuracy (e.g., van der Linden, 2016) or it can be linked to response accuracy models. For example, Klein Entink et al (2009) simultaneously model item accuracy and response times using predictors based on item features, as in LLTM. Molenaar et al (2015) developed a generalized linear framework for modeling person trait differences using the joint impact of accuracy and response time. Also, response time also has been incorporated into modeling within-person differences in response strategies for different items (De Boeck & Jeon, 2019; Molenaar & De Boeck, 2018).

The models described above are just a sample of the many models that have been developed to understand cognitive processing using response accuracy and/or response time data.

Applications of psychometric modeling for cognitive processes

A study on aptitude measurement will be presented to illustrate the potential of applying psychometric modeling to understanding cognitive processes. Although the author has studied fourteen different types of items, including mathematical achievement and reasoning items, as well as paragraph comprehension items, the example below has the more extensive data available that is needed to illustrate implications of the Response Processes to other aspects of validity; namely, Internal Structure and Relationships to Other Variables.

For high stake tests, a single score is often used to make decisions. Thus, Rasch family models are applied to many achievement and aptitude tests in the United States because total score is a sufficient statistic for trait level. Scores weighted by item discrimination can be problematic, as trait level estimates can differ for the same total score. Using Rasch model estimates of trait levels avoids the potential problems and legal challenges that can result from item weighting. Thus, in the current study, only Rasch family models will be used.

In the studies presented below, the Spatial Learning Ability Test (SLAT) is used. SLAT consists of items that involves selecting the three-dimensional display that results from folding a two-dimensional display. The properties and validity of SLAT has been examined in a series of studies (Embretson, 1992, 1996, 2004, 2021a, 2021b; Ivie & Embretson, 2006). Various results on cognitive processes research and its implications for validity and test design are presented below.

A major goal in these studies was to determine if examinees differ in strategies applied in responding to the items. It was hypothesized that at least three different strategies may be involved; (1) a spatial strategy, in which the unfolded object is mentally rotated and folded, (2) a verbal strategy, in which the position of the figure on the non-adjacent side is verbally tracked for position and (3) guessing. Thus, a mixture IRT model (Rost & van Davier, 1995) was applied to identify groups of examinees with varying strategies. Obviously, the spatial strategy is the most construct-relevant strategy and the guessing strategy is construct-irrelevant. The verbal strategy, in contrast, is unclear, especially if spatial tasks in general can be solved by verbal methods.

Methods

SLAT items

Figure 2 presents a SLAT item with unique and directionally sensitive figures on the six sides of a cube that results when folded. Of the three sides show on the folded cubes, at least two sides must be adjacent. The sources of spatial complexity in the various items on SLAT are (1) degrees of rotation, to align the two adjacent sides with the three-dimensional view (0, 90 or 180 degrees) and (2) number of surfaces carried, to view the third side on the folded cube (1, 2, or 3). On Fig. 2, the correct answer is #3, which involves a 90 degree rotation and three surfaces carried. The distractors displayed a fold cubed with a non-attached surface. Side marker objects were varied so that different

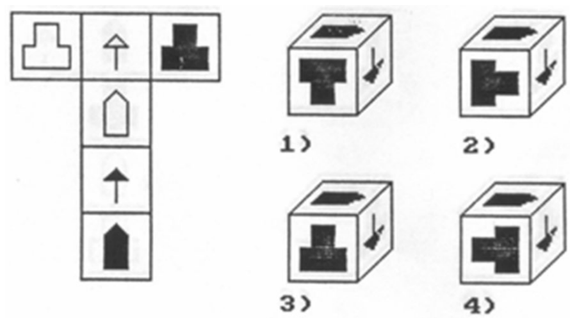

Fig. 2 An item for the spatial learning ability test

items with the same combinations of sources of cognitive complexity (i.e., degrees rotation and surfaces carried) could be produced. An automatic item generator was developed SLAT items.

Test

A SLAT form with 28 items was developed using an automatic item generator. To support the content aspect of validity, the SLAT form was designed to represent all combinations of degrees of rotation and number of surfaces carried.

Examinees

SLAT was computer-administered to a sample of 748 young adults. No time restrictions were imposed.

Models applied

The program winMIRA (von Davier, 2004) was applied to the item accuracy responses to determine if the examinees varied in strategies. With winMIRA, person fit indices are used to determine separate latent classes of examinees that vary in response patterns. The fit indices are derived from the likelihood of an examinees' responses using item parameter estimates from a Rasch model. Separate latent classes, with different patterns of item difficulty, are developed to increase person fit and the log likelihood of the data. Following the identification of latent classes, LLTM and response time models were applied separately within classes to determine the relative impact of the sources of cognitive complexity.

It should be noted that joint modeling procedures for accuracy and response time are available (e.g., Klein Entink et al., 2009), and can include measures of cognitive complexity as a basis to define classes. This method was not applied for several reasons, including a primary interest in classes representing the relative difficulty of items, a possible instability in relationships (i.e., within- person correlations of item difficulty with response time vary widely in this data; Embretson, 2021b) and the anticipated within-class sample sizes may be too small for the required Bayesian estimation method.

Table 1 Fit of successive numbers of latent classes for response accuracy

Number classes	Number parameters	− 2lnL	Chi square difference	BIC	AIC
1	29	24,746.14		24,938.05	24,804.15
2	59	24,529.13	216.96*	24,919.61	24,647.19
3	89	24,397.68	131.50*	24,986.63	24,575.68
4	119	24,320.20	77.48*	25,107.68	24,558.21
5	149	24,257.22	62.98*	25,243.22	24,555.23

*$p < 0.01$

Table 2 Descriptive statistics and internal consistency reliability for person estimates in four-class solution

Class	Size	Trait estimates			Response time	
		Mean	SD	Rel.	Mean	SD
1	0.334	1.192	1.340	0.802	25.281	5.763
2	0.292	0.059	0.700	0.620	25.121	6.700
3	0.201	− 0.846	0.512	0.272	21.493	7.395
4	0.173	1.132	1.082	0.745	25.245	5.424

Results

Table 1 presents results on fit for a successive number of latent classes based on the person by item data. Table 1 shows that as the number of classes increases that the log likelihood index, $-2\ln L$, decreases. The other indices, based on the log likelihoods, show somewhat varying patterns. The BIC index, which typically is a conservative index, decreases only up to three classes. In contrast, the X^2 difference tests are significant up to five classes and the AIC index decreases up to five classes. However, the five-class solution has a class with only four percent of the cases (i.e., 30 examinees), which would be too few for meaningful comparisons with other classes. Further, the AIC decrease is minimal from four to five classes. Thus, the four-class solution was selected.

To provide a comparison, an exploratory multidimensional IRT model was also fit to the data. Although a two-dimensional model had significantly better fit ($\chi^2 = 108.38$, df = 27, $p < 0.01$), the BIC statistic increased from 1 dimension (BIC = 24,963.09) to 2 dimensions (BIC = 25,033.37). Thus, the mixture model has greater support. Comparing the mixture model to a between-item multidimensional model (see Rijmen & de Boeck, 2005) was not attempted as number of item categories would be too high (i.e., 9 categories for number of surfaces by degrees rotation).

Internal structure by class

Table 2 presents descriptive statistics for examinees in the four-class solution, including mean trait levels and mean item response times. It can be seen that class size varies, with Class 1 and Class 2 as the largest classes and Class 3 and Class 4 as smaller classes. The mean trait levels differed significantly between classes (F = 227.466, df = 3, 744, $p < 0.001$), with the highest means for Class 1 and Class 4. Class 2 has a substantially lower mean than Class 1 and Class 4, while Class 3 has a very low mean. The mean item response times also differed significantly between classes (F = 13.929, df = 3, 744,

$p < 0.001$), with Class 3 having a substantially lower mean than the other classes. Finally, internal consistency reliability varies between classes; with Class 1 and Class had moderately high reliabilities, while Class 2 and Class 3 had lower reliabilities.

Response processes: cognitive complexity by class

Table 3 presents the LLTM results on cognitive complexity of item responses with each class. It can be seen that the fit index of the cognitive complexity models is moderately high in three classes, ranging from 0.664 to 0.723. This index is based on log likelihood ratios and is comparable in magnitude to a multiple correlation. Class 3, however, has a lower fit index of 0.481. The predictor weights for the cognitive complexity variables indicate that the Number of Surfaces Carried is significant in all classes, but that its magnitude varies greatly, ranging from 0.223 in Class 3 to 0.931 in Class 2. Degrees of Rotation, however, also varied substantially, ranging from 0.010 to 0.604. Class 4 had the highest weight and strongly significant weight ($p < 0.001$), while Class 2 and 3 had very small weights but significant weights ($p < 0.05$). However, Degrees of Rotation did not have a significant weight in Class 1. Finally, the interaction term was significant in all four classes.

To understand the implications of the predictors, Fig. 3 shows the relationship of item difficulty by Number of Surfaces Carried and Degrees of Rotation for the four classes. It can be seen that for Class 3, the item difficulties varied little by Degrees of Rotation except if Number of Surfaces Carried equals one, where the displayed the surfaces in the response options are adjacent in the item stem. In Class 1, Degrees of Rotation is not related to item difficulty at any value for Number of Surfaces Carried. Class 2 and Class 4 have opposing effects for Number of Surfaces Carried on Degrees of Rotation. That is, for Class 4, the impact of Degrees of Rotation increases directly with Number of Surfaces Carried. In contrast, for Class 2, Degrees of Rotation varies only when Number of Surfaces Carried equals 1 (i.e., when the sides are adjacent).

Table 4 presents the cognitive complexity models of item response times, measured as lnRT. Significant multiple correlations ($p < 0.05$) were observed only for Class 1 and Class 4, while Class 2 had a marginal significance ($p = 0.072$). The weight for Number of Surfaces Carried was significant in Class 1, Class 2 and Class 4. Degrees of Rotation, however, was statistically significant only in Class 4. The interaction term was not a significant predictor in any class.

Relationship to other variables

Trait levels in the various classes were also related to external variables. Test scores were available for the Armed Services Vocational Aptitude Test (ASVAB). Figure 4 shows the significant correlations of SLAT trait scores within each class to three ASVAB scores, Verbal Ability, Mathematical Reasoning and Technical Aptitude. It can be seen that Class 3 SLAT trait scores had no significant relationships while Class 1 SLAT trait scores were significantly related to all three ASVAB scores. Class 2 SLAT trait scores correlated somewhat with Mathematical Reasoning but more strongly with Technical Aptitude. Class 4 SLAT trait scores correlated only with Mathematical Reasoning.

Finally, the relationship of class membership to background variables was examined. Gender and educational level were available for 603 of the 748 examinees. For gender,

Table 3 Cognitive complexity models of item response accuracy by class

	Class 1			Class 2			Class 3			Class 4		
	b	SE	*t*-test	b	SE	*t*-test	b	SE	*t*-test	b	SE	*t*-test
Surfaces carried	0.800	0.044	18.143	0.931	0.041	22.531	0.223	0.049	4.512	0.701	0.062	11.252
Degrees rotation	0.010	0.042	0.240	0.094	0.040	2.700	0.134	0.039	3.044	0.604	0.056	10.770
Interaction	0.141	0.060	20.432	− 0.324	0.054	− .5.884	− 0.231	0.062	− 3.712	0.491	0.083	5.904
Fit index		0.664			0.716			0.481			0.723	

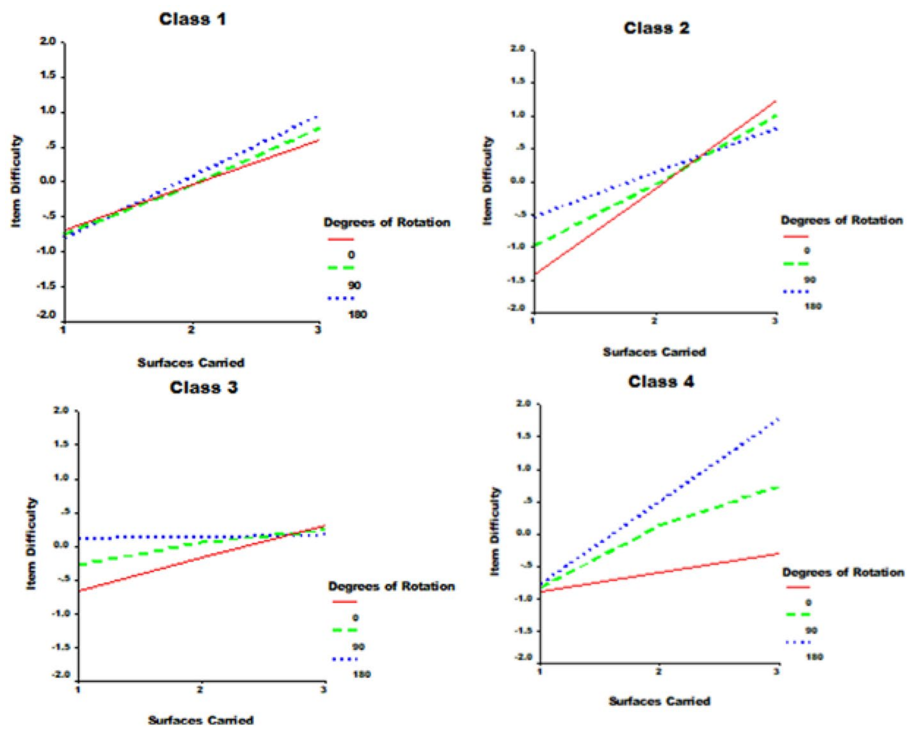

Fig. 3 Item difficulty by cognitive complexity in four latent classes

class membership differed significantly ($X^2 = 10.213$, df = 3, $p = 0.017$) between female and male examinees. Females had approximately 10 percent higher membership in both Class 2 and Class 3 than males. Education level, defined as high school only versus college, did not vary significantly between classes ($X^2 = 2.886$, df = 3, $p = 0.413$).

Discussion

These results strongly suggest the presence of latent classes of examinees, with varying patterns of item difficulty, impact SLAT in several ways. That is, the latent classes vary not only in mean trait levels and response times, but also in the impact of cognitive complexity on item difficulty and item response time, as well as other aspects of test scores such as internal consistency and their external relationships. Thus, three aspects of validity, Response Processes, Internal Structure and Relationship to Other Variables varied across classes.

But what are the specific differences in response processes between the classes? Since the classes vary in item difficulty patterns, they most likely reflect varying strategies involved in item solving. Class 3, for example, most likely involves guessing as a major strategy. With weak or no relationships of item cognitive complexity variables to either accuracy or response time, as well as very low mean trait levels and reliability, and a relatively lower mean item response time it appears that specific item content makes little difference in the responses. Class 1, on the other hand, probably involves a verbal strategy to solve items. That is, examinees in this class are most likely not mentally rotating

Table 4 Cognitive complexity models of item log response times by class

	Class 1			Class 2			Class 3			Class 4		
	b	*t*-test	Prob.	b	*t*-test	Prob.	b	*t*-test	Prob.	b	*t*-test	Prob.
Surfaces carried	0.545	3.226	0.004	0.404	2.247	0.034	−0.059	−0.286	0.777	0.468	2.807	0.010
Degrees rotation	0.260	1.548	0.135	0.304	1.699	0.102	0.047	0.227	0.822	0.408	2.460	0.021
Interaction	−0.039	−0.230	0.820	−0.115	−0.642	0.527	0.014	0.067	0.947	−0.088	−0.528	0.602
Fit index		0.580			0.497			0.083			0.594	

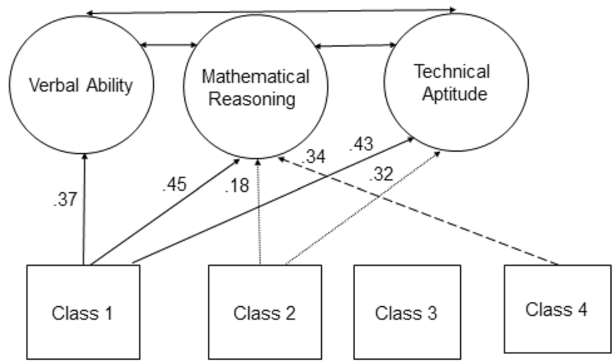

Fig. 4 Correlations of SLAT scores with external test scores by class

items since Degrees Rotation has no relationship to either item accuracy or response time. Class 4, on the other hand, appears to be a spatial strategy, involving mentally folding the stem into the three-dimensional views. That is, both sources of item cognitive complexity are significantly related to response accuracy and response time in this class. Class 2, on the other hand, may involve a combination of strategies. Degrees Rotation did impact response accuracy when the surfaces are adjacent, but not otherwise. Since this class also had a lower mean trait level and a lower reliability, the other strategy may be mostly guessing. In any case, the results on the classes support differences in the Response Processes aspect of validity.

The classes also differed in other aspects of validity. First, the Internal Structure aspect varied. Classes means varied, with Class 1 and Class 4 having high approximately equal means. Class 2 had a moderate mean while Class 3 had a very low mean, indicating guessing. Further, Class 1 and Class 4 had moderately high reliability, indicating consistency across items. Class 2 was substantially lower, indicating some inconsistency in responses and again supporting a combination of item solving strategies. Finally, Class 3 had a very low reliability, again indicating guessing.

The relationship of SLAT scores to the Relationship to Other Variables aspect of validity was also impacted by the latent classes. At the extremes, SLAT scores in Class 1 were significantly related to all three external aptitude measures while SLAT scores in Class 3 were not significantly correlated with any of the aptitude tests. In fact, Class 1 had the only significant correlation of SLAT scores with Verbal Ability, further supporting verbal processing of SLAT items. In contrast, Class 4 SLAT scores were significantly correlated with only with Mathematical Reasoning. Class 2, on the other hand, had a weak correlation with Mathematical Reasoning and a stronger correlation with Technical Aptitude.

Finally, the class in which an examinee had the best fit was related to one of the two background variables that were available. That is, gender varied significantly between the classes, with females more likely to be in the two classes with lower mean trait levels. This finding may have implications for the test Consequences aspect of validity.

Thus, response processes, and its impact on other aspects of validity, by latent classes that are supported as representing differences in item solving strategies. These results suggest that score interpretation, as SLAT measuring a spatial ability, may be somewhat misleading if items can be solved by a more verbal analytic strategy as well. Thus, changes in item design or in test instructions should be explored to determine if a more common strategy in item solving can be implemented. If the verbal strategy remains, SLAT scores should be interpreted as involving aptitude for solving spatial tasks, which may or may not involve spatial processing.

Summary and conclusions

The purpose of this paper was to demonstrate that understanding the cognitive processes, skills and strategies that examinees use in testing have important implications for multiple aspects of validity and score interpretability. Interest in response processes has increased dramatically recently since both item responses and response times are often available with computerized testing. However, although IRT-based modeling of cognitive processes has long been available, it has not often used in testing. Also, many item response time models have been developed but also are not often used in testing.

To increase the potential impact of cognitive process research in testing, this paper presented a brief overview of some IRT-based models that are relevant to understanding response processes. Further, the potential impact of response processes research on the various aspects of validity then was considered in an integrated validity model. It was shown that cognitive processes research in the integrated model can guide both test interpretation and future changes in item and test design.

An example of response processes research on an aptitude test using two IRT-based models was presented. The mixture IRT model was applied to a spatial aptitude test and four latent classes were identified. Cognitive complexity modeling of item difficulty and item response times indicated that the classes varied in problem solving strategies. It was shown how the findings impacted other aspects of validity and had potential implications for score interpretations and test design. For example, the processing differences found between the classes in the current study implies that the test does not necessarily measure "spatial ability" for all examinees. That is, finding a class with apparently a verbal strategy for item solving implies that test scores perhaps should be interpreted as the "ability to solve spatial problems".

Similar modeling procedures could be applied to many other tests to examine impact on score interpretation. Applying mixture modeling procedures to other ability tests may lead to similar interpretation changes if class differences in processes are found. On the other hand, finding latent class differences on achievement tests may not necessarily imply changes in interpretation as performance level is the main concern. Instead, differences between the latent classes could provide useful diagnostic information for subsequent instruction and intervention. More research on other tests is needed to determine such effects and possible advantages.

It should be noted that this paper used the most basic models to understand response processes, including the mixture Rasch model (Rost & van Davier, 1995) and the linear logistic test model (Fischer, 1973). As noted by de Boeck and Jeon (2019), many appropriate models are now available. This includes more integrated models, such as Klein Entink et al. model (2009), models especially devoted to response times (e.g., Molenaar et al., 2015; van der Linden & Fox, 2016), models to analyze specific strategies applied by examinees to each item (Molenaar & de Boeck, 2018) and many more. Hopefully test developers will be motivated to apply the many available models as appropriate so that test validity can be enhanced.

Abbreviations

SLAT	Spatial learning ability test
IRT	Item response theory
LLTM	Linear logistic test model
LPCM	Linear partial credit model

Acknowledgements
No other individuals contributed to this manuscript.

Author contributions
The author contributed uniquely to all aspects of the manuscript. All authors read and approved the final manuscript.

Funding
The current study was not funded.

Availability of data and materials
Data will not be shared as military consent for other users is not available.

Declarations

Ethics approval and consent to participate
The empirical analyses presented in the article is a reanalysis of data collected previously (1992) by the military on the author's test SLAT.

Consent for publication
Not applicable.

Competing interests
There are not competing interests relevant to this study.

Received: 9 June 2022 Accepted: 6 March 2023
Published online: 21 March 2023

References

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

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

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

De Boeck, P., & Wilson, M. (Eds.) (2004) *Explanatory item response models* New York: Springer.

De Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology, 10*, 102.

De Boeck, P., & Wilson, M. (2016). Explanatory item response models. In W. J. van der Linden (Ed.), *Handbook of item response theory. Vol. 1. Models* (Vol. 1, pp. 565–580). Chapman & Hall/CRC.

Embretson, S. E. (2017). An integrative framework for construct validity. In A. Rupp & J. Leighton (Eds.), *The Handbook of Cognition and Assessment*. New York: Wiley-Blackwell.

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

Embretson, S. (1992). Measuring and validating cognitive modifiability: An ability in the spatial domain. *Journal of Educational Measurement, 29*, 25–50.

Embretson, S. E. (1996). Cognitive design systems and the successful performer: A study on spatial ability. *Journal of Educational Measurement, 33*, 29–39.

Embretson, S. E. (2004). Applications of two IRT models for construct validation to issues about spatial ability. *Metodologia De Las Ciencias Del Comportamiento., 5*, 159–180.

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

Embretson, S. E. (2021b). Understanding examinees' item responses through cognitive modeling of response accuracy and response times. Keynote address at the *Theory-based Construction of Process Indicators Conference*. Co-organized by the International Association for the Evaluation of Educational Achievement (IEA), the Leibniz Institute for Research and Information in Education (DIPF), and the Centre for International Student Assessment (ZIB).

Embretson, S. E. (2021a). Response time relationships within examinees: Implications for item response time models. In M. Wiberg, D. Molenaar, J. González, U. Bockenholt & J.-S. Kim (Eds.), *Quantitative psychology.* New York: Springer.

Embretson, S. E., & Yang, X. (2013). A multicomponent latent trait model for diagnosis. *Psychometrika, 78*, 14–36.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359–374.
Fischer, G. H., & Parzer, P. (1991). An extension of the rating scale model with an application to the measurement of change. *Psychometrika, 56*, 637–651.
Ivie, J. & Embretson, S. E. (2006). Spatial learning ability test. In N. Salkind (Ed.), *Encyclopedia of measurement and statistics*. London: Sage.
Janssen, R. (2016). Linear logistic models. In W. van der Linden (Ed.), *Handbook of item response theory: Models, statistics and applications*. New York: Taylor & Francis Inc.
Klein Entink, R. H., Kuhn, J.-T., Hornke, L. F., & Fox, J.-P. (2009). Evaluating cognitive theory: Joint modeling approach using responses and response time. *Psychological Methods, 14*, 54–75.
Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36*, 437–448.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education/Macmillan.
Meulders, M., & Xie, Y. (2004) Person by item predictors. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models. A generalized linear and nonlinear approach* (pp. 213–240). New York: Springer.
Molenaar, D., & De Boeck, P. (2018). Response mixture modeling: Accounting for heterogeneity in item characteristics across response times. *Psychometrika, 83*, 279–297.
Molenaar, D., Terlinckx, F., & van der Mass, H. L. (2015). A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times. *Multivariate Behavioral Research, 50*, 56–74.
Rijmen, F., & De Boeck, P. (2002). The random weights linear logistic test model. *Applied Psychological Measurement, 26*, 271–285.
Rijmen, F., & De Boeck, P. (2005). A relation between a between-item multidimensional IRT model and the mixture Rasch model. *Psychometrika, 70*, 481–496.
Rost, J., & von Davier, M. (1995). Mixture distribution Rasch models. In G. Fischer & I. Molenaar (Eds.), *Rasch models: Foundation, recent developments and applications* (pp. 257–268). Springer.
van der Linden, W. (2016). Lognormal response time model. In W. J. van der Linden. (Ed.). *Handbook of item response theory: Models, statistics and applications*. New York: Taylor & Francis Inc.
van der Linden, W. J., & Fox, J. P. (2016). Joint hierarchical modeling of responses and response times. In W. J. van der Linden (Ed.), *Handbook of item response theory, volume one: Models* (pp. 481–501). CRC Press, Taylor and Francis Group.
von Davier, M., & Rost, J. (1995). Polytomous mixed Rasch models. In G. Fischer & I. Molenaar (Eds.), *Rasch models: Foundation, recent developments and applications* (pp. 371–379). New York: Springer.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.