Knowledge of Solution Strategies and IRT Modeling of Items for Transitive Reasoning

Klaas Sijtsma and Anton C. Verweij

*Applied Psychological Measurement* 1999; 23; 55
DOI: 10.1177/01466219922031194

The online version of this article can be found at:
http://apm.sagepub.com/cgi/content/abstract/23/1/55

Published by:

SAGE Publications
http://www.sagepublications.com

Additional services and information for *Applied Psychological Measurement* can be found at:

Email Alerts: http://apm.sagepub.com/cgi/alerts

Subscriptions: http://apm.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

Citations (this article cites 11 articles hosted on the SAGE Journals Online and HighWire Press platforms):
http://apm.sagepub.com/cgi/content/refs/23/1/55
Componential item response theory (CIRT) is presented as a model-oriented approach to studying processes and strategies underlying the incorrect/correct responses to cognitive test tasks. CIRT is contrasted with a data-oriented approach in which verbal explanations for incorrect/correct responses are collected during the test phase and incorporated in the scoring. Alternatively, the psychologically meaningful data are modeled by unidimensional item response theory models. Verbal explanations for each examinee and task were collected from transitive reasoning tasks in addition to the incorrect/correct responses. Two datasets were compiled, one reflecting the common incorrect/correct scoring and one showing whether a deductive strategy had been used to produce a correct response. The Mokken model of monotone homogeneity, the partial-credit model, and the generalized one-parameter logistic model were used to analyze both polytomous datasets. Results showed that combining knowledge of solution strategies with IRT modeling produced a useful unidimensional scale for transitive reasoning.

Index terms: cognitive strategies, componential IRT models, generalized one-parameter logistic model, Mokken model of monotone homogeneity, partial-credit model, solution strategies, transitive reasoning.

The ability to integrate knowledge about physical relations between objects, i.e., premise information, into a conclusion about an unknown transitive relation is called transitive reasoning and is important in cognitive development (e.g., Piaget & Inhelder, 1941). A transitive reasoning task might consist of three sticks A, B, and C, of different length, denoted Y, such that \( Y_A < Y_B < Y_C \). A transitive inference about the length relation between sticks A and C has been drawn if the examinee concludes that \( Y_A < Y_C \) given his or her knowledge that \( Y_A < Y_B \) and \( Y_B < Y_C \). The conclusion is the result of a deduction from the premise information. Obviously, this conclusion is incorrect or correct and can be coded 0-1, as is common for cognitive test data.

In psychometrics, there is a growing interest in modeling the cognitive processes and solution strategies that underlie the 0-1 scores reflecting the incorrect/correct responses to cognitive tasks such as transitive reasoning tasks. This interest has resulted in the development of componential item response theory (CIRT) modeling (e.g., Embretson, 1985, 1991; Fischer, 1974, 1995; Kelderman & Rijkes, 1994; Mislevy & Verhelst, 1990; Rost, 1996). CIRT modeling of underlying processes provides better understanding of the construct measured by the test, knowledge of processes and strategies that may be useful for remedial teaching, and insight into the construction of a balanced set of items that induce behavior governed by the latent trait of interest. Modeling of cognitive processes can be done by hypothesizing a particular parameter structure for the item parameters or the person parameters, or by decomposing the task into subtasks and modeling this subtask structure.

The alternative approach implemented here concentrated on the data used for studying solution strategies underlying transitive reasoning and building a scale for transitive reasoning using...
item response theory (IRT) models. This study focused on expanding data collection beyond incorrect/correct responses by collecting verbal explanations for responses given. This information was incorporated into the item scores, and the fit of a unidimensional IRT model to the data was tested. By incorporating the strategy-use information into the data, the meaning of item scores was unequivocal and CIRT modeling was unnecessary.

A Model-Oriented Approach: Three CIRT Models

The Linear Logistic Test Model

Let
\[ X_i \] be the random variable for the score on item \( i \) \((i = 1, \ldots, k)\),
\( x \) represent a particular item score, usually reflecting incorrect/correct responses, and
\( \theta \) represent the latent trait.
The linear logistic test model (LLTM; Fischer, 1974, 1995; Scheiblechner, 1972) models the latent item difficulty parameter in the Rasch (1960) model as a linear combination of \( G \) basic parameters \( \eta_g \), with weights \( q_{ig} \), for the difficulty of a task characteristic or a subtask in a solution strategy:

\[
P(X_i = x|\theta) = \frac{\exp \left( x(\theta - \sum_{g=1}^{G} q_{ig} \eta_g - c) \right)}{1 + \exp \left( \theta - \sum_{g=1}^{G} q_{ig} \eta_g - c \right)}, \quad x = 0, 1.
\]

In Equation 1, \( c \) is a normalization constant for the item parameters. The number of basic parameters, \( G \), and the weights \( q_{ig} \), must be known before the model is tested. The basic parameters, \( \eta_g \), are estimated from the data. Butter, De Boeck, & Verhelst (1998) proposed a method to estimate the weights \( q_{ig} \) from the data. This yields an exploratory approach to studying processes and strategies. The person parameters are of little interest in applications of the LLTM. Further, the LLTM assumes that the same cognitive process for each item is used by each examinee (Van Maanen, Been, & Sijtsma, 1989).

The Multidimensional Polytomous Latent Trait Model

The multidimensional polytomous latent trait model (MPLT; Kelderman & Rijkes, 1994; Rijkes, 1996) allows for modeling the use of several strategies in one test and for strategy shift by an examinee. The MPLT model models \( T \) latent person parameters \( \theta_t \) with weights \( B_{ix} \) in a linear combination. It also allows for polychotous items with different numbers of response categories:

\[
P(X_i = x|\theta_1, \ldots, \theta_T) = \frac{\exp \left( \sum_{t=1}^{T} B_{ix} \theta_t + \phi_{ix} \right)}{\sum_{v=0}^{m_i} \exp \left( \sum_{t=1}^{T} B_{ix} \theta_t + \phi_{ix} \right)}, \quad x = 0, \ldots, m_i.
\]

\(\phi_{ix}\) in the MPLT indicates the easiness of arriving at response category \( x \) of item \( i \). To test the MPLT model, the number of person parameters, \( T \), and the weights, \( B_{ix} \), must be specified first. Strategy shift can be incorporated by embedding Equation 2 in a latent class structure (Rijkes, 1996, pp. 87–89).
The Multicomponent Latent Trait Model

The multicomponent latent trait model (MLTM; Embretson, 1985, 1997) decomposes the item into $Q$ subtasks with a separate person parameter for each subtask. The probability of giving the correct response to a subtask is modeled using the Rasch (1960) model. The relation between item performance and performance on the subtasks is multiplicative, so that a noncompensatory model is defined,

$$P(X_i = 1 | \theta_1, \ldots, \theta_Q) = \prod_{q=1}^{Q} P(X_{iq} = 1 | \theta_q) = \prod_{q=1}^{Q} \frac{\exp(\theta_q - \delta_{iq})}{1 + \exp(\theta_q - \delta_{iq})}, \quad (3)$$

where $\delta_{iq}$ is the difficulty of subtask $q$ of item $i$. To estimate the parameters, data must be collected at the subtask level (Embretson, 1985). Thus, the examinee responds to subtasks instead of the entire task. This forces examinees into a strategy and, therefore, choice of strategy is not an issue. Maris (1992) presented an approach that estimates MLTM parameters without subtask data.

Important for the present study was that CIRT approaches impose a formal structure on the incorrect/correct data. This structure reflects a hypothesis about an underlying cognitive process or solution strategy. An assumption of CIRT is that the incorrect/correct responses contain enough information to test hypotheses about the cognitive processes or solution strategies. Whether this is reasonable must be determined by the data.

A Data-Oriented Approach

The 0-1 scores obtained from most cognitive tests reflect only whether the response was incorrect or correct. These responses are often the result of various processes and strategies elicited by particular task and/or person characteristics. For example, Van Maanen et al. (1989) found that their examinees could be divided into four groups that used different solution strategies to solve balance problems (Siegler, 1976). The data-oriented approach to studying cognitive processes and solution strategies relies heavily on verbal protocols from which the strategy that led to the task response is deduced.

First, evidence is collected about the different strategies used by a particular population to solve particular tasks (e.g., Van Maanen et al., 1989). Next, the relation between use of particular strategies and characteristics of the tasks and the presentation procedures of the tasks is studied by experiments. The results determine the task characteristics and the presentation procedures that are the most appropriate for provoking a particular solution strategy. After a set of tasks is designed, responses and verbal explanations for these responses are collected during the test administration phase. A score of 1 denotes that a correct response was produced by an appropriate strategy and a score of 0 denotes all other responses. Because these data reflect whether the solution strategy that led the examinee to a correct response was appropriate, the item scores have a clear meaning.

In this study, two kinds of data were used as the input for an IRT analysis of transitive reasoning tasks: (1) incorrect or correct answers scored 0 or 1 regardless of the strategy used and (2) scores of 0 resulting from no deductive strategy or an incorrect response, or scores of 1 resulting from a correct response based on a deductive strategy. These scores reflected transitive reasoning as deductive inference. The transitive reasoning tasks were based on studies by Verweij (1994; Verweij, Sijtsma, & Koops, in press) that used verbal protocols to determine the task characteristics and presentation procedure most appropriate to provoke a deduction from the premises instead of a different solution strategy. Nine transitive reasoning tasks were used.

Tasks were polytomously scored because tasks consisted of one, two, or three items, each of which was dichotomously scored, and because the nesting of these items within tasks may have
caused some dependence among responses given by the same examinee. The datasets were thus analyzed at the task level using unidimensional IRT models for ordered polytomous scores.

**IRT Models for Polytomous Data Analysis**

Several IRT models for ordered polytomous item scores were used to analyze the transitive reasoning data: (1) The nonparametric Mokken model of monotone homogeneity for polytomous items (MHM; Hemker, Sijtsma, & Molenaar, 1995; Molenaar, 1997) was used to construct an ordinal scale for $\theta$; (2) the partial-credit model (PCM; Masters, 1982), which is a parametric special case of the MHM (Hemker, Sijtsma, Molenaar, & Junker, 1997), was used to investigate whether the items had equal discriminations; (3) the generalized one-parameter logistic model (G-OPLM; Verhelst & Glas, 1995) was used to investigate whether a parametric model could be fitted with varying discriminations across the items, after it had been established that the assumption of equal discriminations was untenable. The G-OPLM is more restrictive than the MHM but less restrictive than the PCM.

**The Mokken Model for Polytomous Items**

The MHM assumes that the $k$ items from the test can be characterized by a single person parameter $\theta$. This is the assumption of unidimensionality. The second assumption is local independence of the task scores. The third assumption is that the response function $P(X_i \geq x | \theta)$ is nondecreasing in $\theta$ (see Hemker et al., 1995). Because this function is not parametrically defined, the MHM is a nonparametric polytomous IRT model. Although it is nonparametric, the MHM is restrictive in the sense that (1) most datasets are not purely unidimensional but are characterized by a dominant latent trait and several less important traits measured by a few or several items, and (2) the response functions are rarely all nondecreasing across the entire scale but deviate slightly from this assumption. The result of a successful MHM data analysis, like most IRT analyses, is usually one or more item sets that predominantly measure one trait with items that have a positive relation with this trait; it does so without disturbing deviations from nondecreasingness that might negatively affect person ordering on the latent trait by means of observable test scores.

An important concept is the item step (Molenaar, 1997). For example, consider a five-object transitive reasoning task with sticks A, B, C, D, and E, and length relations $Y_A < Y_B < Y_C < Y_D < Y_E$. Assume that the premise information is provided by the object pairs AB, BC, CD, and DE, and that the examinee has to solve the length relations within the pairs AC, BD, and CE (Verweij, 1994; Verweij et al., in press); these are three items. Further, consider the scoring rule that involves the strategy information. For this task, three imaginary item steps can be taken. Solving at least $x$ items by deductive reasoning is equivalent to taking at least $x$ item steps. This yields task scores of 0, 1, 2, or 3. Failure to solve any of the items by deductive reasoning represents the lowest ability level for this task. The more items that are solved correctly, the higher the ability level for the item. An item with four score categories is characterized by four item step response functions (ISRFs), $P(X_i \geq x | \theta)$, $x = 0, 1, 2, 3$. However, $P(X_i \geq 0 | \theta) = 1$ is trivial.

Important questions in scale analysis are whether this scoring procedure is justified and whether the sum score on all nine tasks can be used as a transitive reasoning measure. The MHM can be used to answer these questions without assuming a particular parametric form for the ISRF.

Unidimensionality, local independence, and nondecreasingness of the ISRFs imply that all covariances between the tasks must be non-negative (Hemker et al., 1995; Holland & Rosenbaum, 1986). Two methods are available to investigate whether the ISRFs are nondecreasing. First, the scalability coefficient $H$ (Hemker et al., 1995; Molenaar, 1991, 1997) can be used. Let $\sigma_{ij}$ be the covariance between tasks $i$ and $j$ and $\sigma_{ij}^{\text{max}}$ be the maximum possible covariance given the marginals of the bivariate cross-tabulation of the scores on items $i$ and $j$. For $k$ items, coefficient
$H$ is defined as

$$H = \frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \sigma_{ij}}{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \sigma_{ij(\text{max})}}.$$  \hspace{1cm} (4)

Coefficient $H_i$ (Hemker et al., 1995; Molenaar, 1997) indicates whether item $i$ is scalable in accordance with the MHM given the other items used and is defined as

$$H_i = \frac{\sum_{j \neq i} \sigma_{ij}}{\sum_{j \neq i} \sigma_{ij(\text{max})}}.$$  \hspace{1cm} (5)

It can be shown (Hemker et al., 1995; Mokken & Lewis, 1982) that $H \geq \min(H_i)$. Given the MHM, $0 \leq H \leq 1$ and $0 \leq H_i \leq 1$ holds for all $i$ (Hemker et al., 1995). Because positive $H$ values close to 0 provide little information about the MHM (Mokken, Lewis, & Sijtsma, 1986), Hemker et al. (1995) provided guidelines for the use of $H$ for polytomous items. A set of items is considered unscalable if $H < .3$. Weak scalability is when $.3 \leq H < .4$, moderate scalability if $.4 \leq H < .5$, and strong scalability if $.5 \leq H \leq 1.0$. In general, a higher $H$ value implies higher confidence in the person ordering using the total score as an estimate of the latent or true ordering (Mokken et al., 1986).

Second, the empirical regression

$$P(X_i \geq x | R) = \frac{\sum_{j \neq i} n_{r,xi}}{n_r},$$  \hspace{1cm} (6)

was used to approximate the ISRF (Molenaar, 1997).

Let $n_r$ be the size of the group with $R = r$, $n_{r,xi}$ the number of respondents with restscore $R = r$ and score $x$ on item $i$, and $m_i + 1$ the number of response categories of item $i$.

Then

$$\hat{p}_{x|r} = \hat{p}(X_i \geq x | R = r) = \frac{\sum_{j \neq i} n_{r,xi}}{n_r}.$$  \hspace{1cm} (7)

For a particular item and item score, these proportions should be nondecreasing in $R$ except for sample fluctuations (Molenaar, 1997). In case of reversals, the null hypothesis of equal proportions ($\pi_{x|r} = \pi_{x|r+\Delta}$; $\Delta \in N^+$) is tested against the alternative that the second proportion is smaller than the first, using an accurate normal approximation of the hypergeometric distribution (Molenaar, 1970, chap. 4, Formula 2.37).

**The Partial-Credit Model and Generalized One-Parameter Logistic Model**

*The PCM.* The PCM (Masters, 1982) is a parametric IRT model for polytomous item scores. The category response function (CRF) is defined as
where \( \delta_{ij} \) is the location parameter of category \( j \) of item \( i \), and \( \sum_{j=1}^{0} (\theta - \delta_{ij}) \equiv 0 \) for notational convenience. Because the PCM is a special parametric case of the MHM (Hemker et al., 1997), a fitting PCM implies fit of the MHM, although the reverse may not be true.

Even though the MHM fits a dataset, it might be interesting to investigate whether a more restrictive and more informative IRT model also fits the data. Therefore, the scales found by the MHM were also analyzed by means of the PCM.

Fit of the PCM was evaluated using an overall \( \chi^2 \) test \( R_{1c} \) (Verhelst & Glas, 1995) for the null hypothesis that the slopes of the CRFs are equal (see Equation 8; the slope parameter equals 1) within and across all \( k \) items. Fit was also evaluated using one \( \chi^2 \) test (denoted \( \chi^2 \)) and three standard normal tests (denoted \( M_1 \), \( M_2 \), and \( M_3 \)), for each item to test the null hypothesis that expected CRFs equal observed CRFs.

All tests are based on a subgrouping of the examinees based on the total score \( S \) on all \( k \) items. Different tests use different subgroupings. Let \( S^* \) denote the subgrouping variable. A subgroup may contain examinees with different but adjacent total scores \( S = s \). Further, let \( X_i^D = x^D = 0, 1 \) denote a particular dichotomization of the item score of item \( i \) based on the joining of adjacent categories. Finally, let \( n_{s^*} \) be the size of the subgroup defined by \( S^* = s^* \), and \( n_{s^*} \) the number of persons that have a subgrouping score \( S^* = s^* \), and a dichotomized item score \( X_i^D = 1 \), then

\[
\hat{\tau}_{1i|s^*} = \hat{\nu}(X_i^D = 1|S^* = s^*) = \frac{n_{s^*} + 1}{n_{s^*}}.
\]

In addition, a model-based probability, \( \pi_{1i|s^*} = P(X_i^D = 1|S^* = s^*) \), is calculated using the estimated item parameters. For each dichotomization, each test compares differences of observed and expected probabilities

\[
\hat{\pi}_{s^*} - \pi_{s^*} ,
\]

across all subgroups defined by \( S^* \). With \( m_i + 1 \) response categories, \( m_i \) dichotomizations are considered for each item.

A significant item \( \chi^2 \) means that the observed CRFs are different from the expectation given the PCM. Significant positive values of \( M_1 \), \( M_2 \), and \( M_3 \) are indicative of flatter slopes than expected; significant negative values are indicative of steeper slopes. \( M_1 \), \( M_2 \), and \( M_3 \) are the same test based on different subgroupings of the sample.

Item test results are not independent for different items. A deviating item may influence the test results of other items. Because they are based on different subgroupings, \( M_1 \), \( M_2 \), and \( M_3 \) have different power to detect deviations from the model. Occasionally, they may have different signs, which complicates the interpretation. Because of the large number of tests, the significance level must be adapted to protect against chance capitalization.

**G-OPLM.** The third model used was the G-OPLM (Verhelst & Glas, 1995). This model is more restrictive than the MHM and less restrictive than the PCM. The G-OPLM allows a slope index \( a_i \in N^+ \)
to be included in Equation 8, replacing \((\theta - \delta_{ij})\) with \(a_i(\theta - \delta_{ij})\). The results from the \(\chi^2\) tests and the \(M_1, M_2,\) and \(M_3\) item tests for the PCM can be used to hypothesize values for the \(a_i\)s to be inserted in Equation 8 for items with CRFs that are too steep or too flat. Note that the G-OPLM differs from Muraki’s (1992) generalized PCM that has slope parameters \(\alpha_i\), yielding exponents \(\alpha_i (\theta - \delta_{ij})\), rather than indices \(a_i\). The parameters \(\alpha_i\) must be estimated from the data rather than entered by the researcher.

Method

Examinees

The examinees were 417 second-, third-, and fourth-grade students. The mean ages in months were 100.6 (\(n = 139\)), 114.4 (\(n = 140\)), and 122.2 (\(n = 138\)), respectively. The standard deviations were 5.8, 5.4, and 5.2, respectively. Both genders were approximately equally represented in each grade.

Transitivity Tasks

Because the literature (e.g., Brainerd & Kingma, 1984; Brainerd & Reyna, 1990; Sternberg, 1980; Trabasso, 1977) does not provide clear guidelines for constructing a psychometrically sound measurement instrument for transitive reasoning, Verweij (1994; Verweij et al., in press) studied the preference for solution strategies as a function of task format and presentation procedure. To elicit responses based on deduction from the premise information, they found that (1) the relations between objects within a task should be unequal (e.g., sticks should differ in length, if length is the property of interest); (2) length differences between objects containing premise information should be small; (3) premise information should be presented successively (i.e., one at a time; when the items are tested, the inference is drawn while all objects are visually present and arranged in a random order); and (4) verbal explanations should be required to evaluate whether a response was based on deduction from the premise information. When these conditions were not satisfied, stimulus characteristics tended to elicit solution strategies that are not typical of transitive reasoning (Verweij, 1994; Verweij et al., in press).

These results were incorporated in this study. Length, weight, and size were included in the set of tasks because they are useful for the measurement of transitive reasoning (Verweij, Sijtsma, & Koops, 1996) and because a set of physical properties was desired that were representative of transitive reasoning research in general. Tasks consisted of three, four, or five objects. The combination of three physical properties and three numbers of objects led to nine tasks (see Table 1).

<table>
<thead>
<tr>
<th>Table 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description of Transitive Reasoning Tasks</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of objects/label</th>
<th>Material</th>
<th>Common measure</th>
<th>Mean</th>
<th>Difference between adjacent objects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length</td>
<td>3/Le1, 4/Le2, 5/Le3</td>
<td>round, wooden sticks</td>
<td>diameter, .6 cm.</td>
<td>.2 cm.</td>
</tr>
<tr>
<td>Size</td>
<td>3/Si1, 4/Si2, 5/Si3</td>
<td>round, wooden disks</td>
<td>thickness, .4 cm.</td>
<td>.2 cm.</td>
</tr>
<tr>
<td>Weight</td>
<td>3/We1, 4/We2, 5/We3</td>
<td>clay balls</td>
<td>diameter, 5.0 cm.</td>
<td>3.0 gr.</td>
</tr>
</tbody>
</table>

For the size tasks, it was possible that examinees responded to area instead of diameter. Area is a squared measure and, as a result, possible use of a visual strategy (i.e., an examinee concludes that disk A is smaller than disk C because he/she can see that) cannot be excluded. The possible use of a visual strategy was excluded for the weight tasks, because all objects had the same magnitude. The objects were colored white, red, yellow, green, and blue. Colors were randomly determined and
each object had a different color. The objects were identified by color in the conversation between
the experimenter and the examinee.

Administration Procedure
Successive presentation was used, and examinees were individually tested. A three-object task
consisted of two premises, AB and BC, and one item, AC. A four-object task consisted of three
premises, AB, BC, and CD, and two items, AC and BD; and a five-object task consisted of four
premises, AB, BC, CD, and DE, and three items, AC, BD, and CE. The pairs AD (four-object
task) and AD, BE, and AE (five-object task) were not administered because the greater difference
between the objects might have elicited a visual strategy (Verweij et al., in press). For each examinee
the presentation order of the tasks, the premises, and the items was random. Explanations for each
response were recorded to evaluate strategy use.

Item and Task Scoring
The first scoring rule produced 0-1 scores reflecting incorrect/correct responses, respectively.
Because the items for four- and five-object tasks were nested within tasks, it is possible that these
item scores were locally dependent within tasks (e.g., Hambleton & Swaminathan, 1985, pp. 22–
25). Therefore, tasks were used as the units of analysis. A three-object task (one item) was scored
0-1; a four-object task (two items) was scored 0-1-2; and a five-object task (three items) was scored
0-1-2-3. This polytomous dataset was denoted NOSTRA T.

The second scoring rule produced an item score of 1 if a correct response was supported by a
deductive explanation, verbal or nonverbal (e.g., pointing to an object), and a score of 0 otherwise.
For example, if for a five-object task an examinee solved one item correctly using a deductive
strategy, one item correctly using a visual strategy, and one item incorrectly without an explanation,
the task score was 1. In general, task scores were 0-1 (three-object tasks), 0-1-2 (four-object tasks),
and 0-1-2-3 (five-object tasks). This polytomous dataset was denoted DEDSTRAT.

Analysis
To evaluate the fit of the MHM, the data were analyzed with the computer program MSP 3.04
(Molenaar, Debets, Sijtsma, & Hemker, 1994). MSP contains an automated bottom-up item selection
procedure that attempts to construct scales that are in accordance with the MHM (see Hemker et al.,
1995). Given a preliminary selection of items for the first scale, item \( f \) is then selected, which (1)
has positive covariances with each of the selected items; (2) has an \( H_f \) with the selected items of
at least \( c > 0 \); and (3) maximizes the common \( H \) of the selected items including item \( f \), given
all possible choices. If no items are left that satisfy all three conditions, MSP attempts to construct
a second scale from the remaining items. This continues until no item remains unselected or no
additional scales can be constructed. Because a large number of significance tests \( H_i = 0 \) against
\( H > 0 \) is performed at each step of item selection, a progressive Bonferroni
correction protects against chance capitalization across the steps.

To evaluate the fit of the PCM and the G-OPLM, the data were also analyzed with the program
OPLM (Verhelst, 1992). Unlike MSP, OPLM does not select items into subscales. The researcher
uses fit statistics \( (R_{1c}, \chi^2, M_1, M_2, \text{ and } M_3) \) to evaluate the quality of the scale and the items.

Results
Analysis of Verbal Explanations
For each physical property (length, size, and weight), the three-object task had one item, the
four-object task had two items, and the five-object task had three items, resulting in six items for
each property. The mean proportions of examinees who correctly responded to the six items were
.70, .92, and .62, for length, size, and weight, respectively. Most correct responses on length and weight items were based on deductive reasoning (Table 2).

<table>
<thead>
<tr>
<th>Property</th>
<th>Correct Deductive</th>
<th>Correct Visual</th>
<th>Incorrect Deductive</th>
<th>Incorrect Visual</th>
<th>Incorrect None</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length</td>
<td>.51</td>
<td>.12</td>
<td>.18</td>
<td>.09</td>
<td>.10</td>
</tr>
<tr>
<td>Weight</td>
<td>.40</td>
<td>.01</td>
<td>.13</td>
<td>.01</td>
<td>.44</td>
</tr>
<tr>
<td>Size</td>
<td>.21</td>
<td>.68</td>
<td>.04</td>
<td>.01</td>
<td>.05</td>
</tr>
</tbody>
</table>

Table 3

Task Pairs with Negative Covariances for NOSTRAT

Size (Si) With Length (Le) or Weight (We)
  Si1: Le1, We2, We3
  Si2: Le1, Le2, Le3, We1, We2, We3
  Si3: Le3, We1, We2, We3

Following a strategy for finding unidimensional scales suggested by Hemker et al. (1995), several lower bounds $H = c$ were used to select scales conforming to the MHM (Table 4). For $c = 0.0$, .3, .4, .5, and .55, MSP first selected the six length (Le) and weight (We) tasks in one scale ($H = .69$) and then the three size (Si) tasks in another scale ($H = .59$). For $c = .6$, one of the size (Si3) tasks was excluded from the second scale because $H_{ij}$ with respect to the other two size tasks was .58; it should have been at least .6 for inclusion. For $c = .65$, one of the weight (We2) tasks was excluded from the first scale ($H_{ij} = .64$). A second scale could not be formed because none of the remaining task pairs had an $H_{ij}$ that exceeded .65. Given this pattern of results, Hemker et al. (1995) recommended acceptance of the two-dimensional solution for $c = .3$ (Table 5).

Nondecreasingness of the $P(X_i \geq x|R)$ (Equation 6) was investigated by inspection of the observed regressions $\pi_{x|R}$ (see Equation 7). Eighteen regressions were inspected. The minimum allowed restscore group size was 20. Only the first restscore regression of We2 [$P(X_i \geq 1|R)$] had one significant violation (size .07). No violations were detected for Scale 2. Note, however, that the restscore for Scale 2 was based on two items. Thus the number of restscore groups was limited and an accurate analysis of the restscore regressions was not possible.
Table 4
Item Selection by MSP for Different Lower Bounds $H = c$ for NOSTRA T and DEDSTRAT

<table>
<thead>
<tr>
<th>c</th>
<th>Scale 1</th>
<th>Scale 1</th>
<th>Scale 1</th>
<th>Scale 2</th>
<th>Scale 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0 - .55</td>
<td>Le1 Le1 Le1</td>
<td>Le1 Le1</td>
<td>Le1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.6</td>
<td>Le2 Le3 Le2</td>
<td>Le2 Le3</td>
<td>Le3</td>
<td>Le3</td>
<td></td>
</tr>
<tr>
<td>.65</td>
<td>We1 We1 We1</td>
<td>We1 We1</td>
<td>We1 We1</td>
<td>We2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>We3 We3 We3</td>
<td>We3 We3</td>
<td>We3 Si1 Si1</td>
<td>Si1 Si2 Si2</td>
<td></td>
</tr>
</tbody>
</table>

Scale 2

<table>
<thead>
<tr>
<th></th>
<th>Si1</th>
<th>Si2</th>
<th>Si3</th>
</tr>
</thead>
</table>

No Deductive Strategy or Incorrect Response (0) or Deductive Strategy and Correct (1) Data: DEDSTRAT. All 36 covariances between the nine tasks were positive, thus supporting fit of the MHM to these data. For $c = .0, .3, .4, .5, .55, .6, .65$, MSP selected all nine tasks into the same scale (Table 4). For $c = .7$ and .75, seven tasks formed one scale; Si3 ($H_i = .66$) and Le2 ($H_i = .70$) were excluded. For $c = .8$, two scales of three tasks each were formed, and one scale of two tasks; Le2 ($H_i = .75$) was excluded. Given this pattern of results, and recommendations by Hemker et al. (1995), the unidimensional solution for $c = .3$ was accepted (Table 5). Only the second restscore regression of Le2 $P(X_i \geq 2|R)$ had one significant violation: its size was .14.

PCM and G-OPLM Analyses

Incorrect-Correct (0-1) Data: NOSTRA T. For all nine tasks taken together, the overall test led to rejection of the PCM: $R_{1c} = 720$, degrees of freedom (DF) = 40, and $p = 0.00$. The MHM length and weight scale was also rejected: $R_{1c} = 179$, DF = 26, and $p = 0.00$. Because of the large

Table 5
Scalability Coefficients $H_i$ and $H$ for Final Scales

<table>
<thead>
<tr>
<th>Scale and Task</th>
<th>NOSTRA T</th>
<th>DEDSTRAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scale 1</td>
<td>$H_i$</td>
<td>$H$</td>
</tr>
<tr>
<td>Le1</td>
<td>.77</td>
<td>.82</td>
</tr>
<tr>
<td>Le2</td>
<td>.73</td>
<td>.67</td>
</tr>
<tr>
<td>Le3</td>
<td>.69</td>
<td>.62</td>
</tr>
<tr>
<td>We1</td>
<td>.66</td>
<td>.81</td>
</tr>
<tr>
<td>We2</td>
<td>.64</td>
<td>.78</td>
</tr>
<tr>
<td>We3</td>
<td>.66 .69</td>
<td>.63</td>
</tr>
<tr>
<td>Scale 2</td>
<td>$H_i$</td>
<td>$H$</td>
</tr>
<tr>
<td>Si1</td>
<td>.61</td>
<td>.84</td>
</tr>
<tr>
<td>Si2</td>
<td>.59</td>
<td>.78</td>
</tr>
<tr>
<td>Si3</td>
<td>.58 .59</td>
<td>.78 .75</td>
</tr>
</tbody>
</table>
number of significance tests, the item test results (Table 6) were evaluated at a .001 significance level (two-tailed standard normal tests, $z_{\text{crit}} = 3.3$).

Table 6
Item Fit Results of the PCM for Scales From Different Task Dichotomizations (Dichot)
Using NOSTRAT and DEDSTRAT Data

<table>
<thead>
<tr>
<th>Task</th>
<th>Dichot</th>
<th>$\chi^2$</th>
<th>DF</th>
<th>$p$</th>
<th>$M_2$</th>
<th>$M_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOSTRAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Le1</td>
<td>0,1</td>
<td>11.8</td>
<td>2</td>
<td>.003</td>
<td>-3.4</td>
<td>-2.2</td>
</tr>
<tr>
<td>Le2</td>
<td>0,1-2</td>
<td></td>
<td></td>
<td></td>
<td>-3.0</td>
<td>-2.2</td>
</tr>
<tr>
<td></td>
<td>0,1-2</td>
<td>13.3</td>
<td>2</td>
<td>.001</td>
<td>-3.3</td>
<td>.5</td>
</tr>
<tr>
<td>Le3</td>
<td>0,1-2</td>
<td>9.3</td>
<td>2</td>
<td>.009</td>
<td>-3.2</td>
<td>1.3</td>
</tr>
<tr>
<td></td>
<td>0,1-2</td>
<td>3.7</td>
<td></td>
<td>.053</td>
<td>-5.3</td>
<td>1.1</td>
</tr>
<tr>
<td>We1</td>
<td>0,1</td>
<td>6.5</td>
<td>1</td>
<td>.011</td>
<td>-3.3</td>
<td>1.9</td>
</tr>
<tr>
<td>We2</td>
<td>0,1-2</td>
<td></td>
<td></td>
<td></td>
<td>-3.4</td>
<td>-3.4</td>
</tr>
<tr>
<td></td>
<td>0,1-2</td>
<td>13.0</td>
<td>2</td>
<td>.002</td>
<td>-1.3</td>
<td>3.7</td>
</tr>
<tr>
<td>We3</td>
<td>0,1-3</td>
<td></td>
<td></td>
<td></td>
<td>-2.0</td>
<td>7.4</td>
</tr>
<tr>
<td></td>
<td>0,1-2</td>
<td>3.0</td>
<td>2</td>
<td>.222</td>
<td>-1.6</td>
<td>-1.6</td>
</tr>
<tr>
<td></td>
<td>0-2,3</td>
<td></td>
<td></td>
<td></td>
<td>-1.8</td>
<td>3.4</td>
</tr>
<tr>
<td>DEDSTRAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Le1</td>
<td>0,1</td>
<td>3.6</td>
<td>2</td>
<td>.161</td>
<td>-3.5</td>
<td>-3.5</td>
</tr>
<tr>
<td>Le2</td>
<td>0,1-2</td>
<td></td>
<td></td>
<td></td>
<td>-1.3</td>
<td>-1.2</td>
</tr>
<tr>
<td></td>
<td>0,1-2</td>
<td>1.1</td>
<td>2</td>
<td>.299</td>
<td>.0</td>
<td>6.7</td>
</tr>
<tr>
<td>Le3</td>
<td>0,1</td>
<td>17.0</td>
<td>2</td>
<td>.000</td>
<td>3.9</td>
<td>4.8</td>
</tr>
<tr>
<td>We1</td>
<td>0,1</td>
<td></td>
<td></td>
<td></td>
<td>.1</td>
<td>2.0</td>
</tr>
<tr>
<td>We2</td>
<td>0,1-2</td>
<td>8.6</td>
<td>1</td>
<td>.003</td>
<td>-2.8</td>
<td>-1.3</td>
</tr>
<tr>
<td></td>
<td>0,1-2</td>
<td></td>
<td></td>
<td></td>
<td>-3.6</td>
<td>-3.6</td>
</tr>
<tr>
<td>We3</td>
<td>0,1-2</td>
<td>9.9</td>
<td>2</td>
<td>.002</td>
<td>-3.1</td>
<td>-2.6</td>
</tr>
<tr>
<td></td>
<td>0,1-2</td>
<td></td>
<td></td>
<td></td>
<td>-9.3</td>
<td>-2.6</td>
</tr>
<tr>
<td></td>
<td>0-1,3</td>
<td></td>
<td></td>
<td></td>
<td>-2.8</td>
<td>.1</td>
</tr>
<tr>
<td>Si1</td>
<td>0,1</td>
<td>.0</td>
<td>1</td>
<td>.857</td>
<td>-9.3</td>
<td>1.0</td>
</tr>
<tr>
<td>Si2</td>
<td>0,1</td>
<td>8.0</td>
<td>2</td>
<td>.019</td>
<td>-2.8</td>
<td>-2.8</td>
</tr>
<tr>
<td></td>
<td>0-1,3</td>
<td></td>
<td></td>
<td></td>
<td>-2.8</td>
<td>.4</td>
</tr>
</tbody>
</table>

- = Calculation not possible due to dichotomization yielding one low frequency score category.

aLe3 was rescored 0, 1, 2 because original score category 0 was empty.
bScore categories 0 and 1 joined ($n_0 = 3$) and score categories 2 and 3 joined ($n_3 = 5$).
cScore categories 2 and 3 joined ($n_3 = 1$).
d0 frequency in score category 2.
e0 frequencies in score categories 2 and 3.

The high positive standard normal deviates $M_2$ and $M_3$ in Table 6 suggest that We2 and We3 had flatter slopes than expected. This result was found for only one dichotomization of the item scores and this may weaken the conclusion about the slopes. Moreover, because many tasks produced $z$ values close to and, in a few cases, larger than $z_{\text{crit}}$, it appears that most tasks contributed to the overall misfit of the PCM.

Given the indications that the slopes of the CRFs of We2 and We3 may have been flatter than those of the other four tasks, the G-OPLM was fit with imputed slope indices having the following ratios for the subsets of four and two tasks (We2, We3), respectively: 2:1; 3:1, 4:1; and 5:1.
result for the ratio 3:1 was the best, but still led to rejection of the G-OPLM: $R_{1c} = 88$, $DF = 28$, and $p = 0.00$. No important deviations were found at the task level. The program OPLM also suggested a possibly successful set of slope indices: 5 (Le1), 3 (Le2, Le3, We1, We2), and 2 (We3) that resulted in $R_{1c} = 116$, $DF = 25$, and $p = 0.00$ and significant item fit statistics for We2. Thus the G-OPLM could not be fit to these data.

The PCM was not rejected for the MHM size scale: $R_{1c} = .5$, $DF = 2$, and $p = .78$. The test, however, may not have been very powerful because the number of tasks was small and a few response categories had low frequencies. Because of these low frequencies, several item test statistics could not be calculated.

No Deductive Strategy or Incorrect Response (0) or Deductive Strategy and Correct (1) Data: DEDSTRAT. The PCM did not fit the nine tasks combined into the MHM scale of transitive reasoning: $R_{1c} = 193$, $DF = 30$, and $p = 0.00$. The item $\chi^2$ and standard normal deviates (significance level .001, two-tailed standard normal tests, $z_{crit} = 3.3$) suggested that Le3 had flatter slopes than expected (Table 6). The other item tests did not show any significant deviations.

Based on these results, G-OPLM was fit to the data with a small slope index for Le3 relative to the slopes of the other eight tasks. For the slope ratios 2:1 and 3:1, OPLM ran into computational problems; consequently, the following slopes suggested by the program were: 2 (Le2, Le3), 3 (Si1, Si3, We1, We2, We3), and 4 (Le1, Si2). This led to $R_{1c} = 59$, $DF = 31$, and $p = .0017$, and no clearly deviating item fits.

Discussion

The use of the knowledge of solution strategies to score the responses on the transitivity items (DEDSTRAT data) led to a strong MHM scale for the measurement of transitive reasoning by means of deductive inference. When all correct responses were scored 1, two MHM scales were found reflecting the multidimensionality due to use of different solution strategies (see Table 2; the deductive strategy was dominant for length and weight tasks, and the visual strategy was dominant for size tasks). Without knowledge of solution strategies, interpretation of these scales would be more speculative, perhaps inferring the existence of two abilities of transitive reasoning, and thus not revealing the unidimensional scale for transitive reasoning by deductive inference.

The straightforward results for the DEDSTRAT data were based on previous research in which the use of solution strategies was related to presentation procedure and task characteristics. Knowledge of solution strategies is useful for IRT modeling of item response data. These strategies may not only be related to characteristics of the tasks and the testing environment, as was done here, but also to person characteristics.

The misfit of the parametric PCM perhaps can be attributed to the requirement that the slopes of the CRFs are equal within and across all items. Item analyses using the G-OPLM, which allows variation of slopes across (but not within) items, led to better fit results than those using the PCM, but not to fitting models. This result suggests that the rejection of the models also may have been caused by different slopes of CRFs within items and/or by data features other than unequal slopes that disagree with the PCM and G-OPLM assumptions, such as small violations of unidimensionality or local independence. The MHM only requires nondecreasingness of the ISRFs, and thus allows for varying slopes within and across items. It thus appears that a nonparametric model that does not restrict the ISRFs (and, thus, the CRFs) to be fixed by a particular parametric choice was more appropriate to analyze the data.

CIRT modeling is useful if a researcher does not collect verbal explanations for each examinee and each item, as is common in most research. If verbal explanations are collected, it may also be interesting to test whether a CIRT model fits the data. CIRT is useful if a substantive theory is
available that accurately predicts the number ($G$) of basic parameters ($\eta_g$) in the LLTM or the number ($T$) of person parameters ($\theta_i$) in the MPLT model and the weights $q$ and $B$ that must be specified in both models (Equations 1 and 2). The MLTM approach forces examinees into a particular strategy, but cognitive processes relevant to the task structure must be known beforehand. This holds also for a generalized version of the MLTM (Embretson, 1997), the general component latent trait model (GLTM), that imposes a LLTM-like structure on the subtask difficulties $\delta_{iq}$ (Equation 3). Without a theoretical basis for these specifications in each of the CIRT models, testing CIRT models becomes risky, eliciting results based on chance capitalization.

CIRT modeling and the data-oriented approach are different means to accomplish the same objective. A possible disadvantage of CIRT is its heavy reliance on substantive theories that are often not accurate enough to specify a fitting model. A disadvantage of the data-oriented approach is that it is time-consuming; test administration is individual and probing, and recording explanations is time consuming. However, combined with experimentation, the data-oriented approach appears to be very useful because it provides insight into the solution strategies used and, therefore, leads to good scales with clear interpretations.

References


**Author’s Address**

Send requests for reprints or further information to Klaas Sijtsma, Department of Research Methodology (MTO), FSW, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. Email: k.sijtsma@kub.nl.