The Random Weights Linear Logistic Test Model

Frank Rijmen and Paul de Boeck
K. U. Leuven, Belgium

A generalization of the linear logistic test model of G. H. Fischer (1973), the random weights linear logistic test model, is presented. The generalization consists of a random coefficient contribution of item stimulus features to the item difficulties, with the coefficients varying over persons. Whereas in the common linear logistic test model, only the intercept (ability) is considered random over persons, in the random weights linear logistic test model, also some or all of the item stimulus features are considered as having random coefficients. It turns out that the random weights linear logistic test model is a special case of the multidimensional random coefficient multinomial logit model of Adams, Wilson, and Wang (1997). The model is applied to a deductive reasoning task. *Index terms: item response theory, linear logistic test model, multidimensional models, multilevel models, random coefficients models.*

If the aim is to model a cognitive theory taking into account individual differences, one can rely on a variety of item response theory (IRT) models (e.g., Adams, Wilson, & Wang, 1997; Embretson, 1984, 1985; Fischer, 1973; Kelderman & Rijkes, 1994; Maris, 1995). In all these models, item stimulus features with an effect on the cognitive processes as stipulated in the theory are incorporated in the mathematical model, so that the fit of the resulting model can then be used to evaluate the cognitive theory (Embretson, 1998).

To build a mathematical model for a cognitive theory, one has not only to identify the items' relevant stimulus features, but one also must specify how these stimulus features (together with person characteristics) are supposed to determine the probability that an item is solved correctly. That is, one should specify the function that relates the item features with the probability that an item is solved correctly. With respect to the latter, different psychometric models exemplify different approaches. One may distinguish between two broad types of approaches. In the first approach, as illustrated in the linear logistic test model (LLTM; Fischer, 1973), item features are seen as sources of item difficulty. In the second, item features are seen as the basis of individual differences, as in the multidimensional models of Embretson (1984) and Kelderman and Rijkes (1994).

The multidimensional random coefficients multinomial logit model (MRCMLM; Adams, Wilson, & Wang, 1997) is a general IRT model subsuming many others. The model offers a general framework that can handle dichotomous as well as polytomous responses. It subsumes unidimensional IRT models, between-item multidimensional models (in which each item only measures one dimension, e.g., when a test consists of different subscales), and within-item multidimensional models (in which each item can measure more than one dimension). Furthermore, rater effects and multifaceted data formats in general (Linacre, 1989/1994) can also be handled.

In the context of modeling a cognitive theory, its generality allows for incorporating the two approaches just discussed. For dichotomous items, the probability that a person will give a correct answer on item i is, conditional on his ability vector $\boldsymbol{\theta}_{\mathbf{n}}$:

Applied Psychological Measurement, Vol. 26 No. 3, September 2002, 271–285
© 2002 Sage Publications

Downloaded from http://apm.sagepub.com at Lmog Ku Leuven Campusbiblio on April 11, 2008
© 2002 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

$$P(X_{in} = 1 \mid \mathbf{A}, \mathbf{B}, \boldsymbol{\eta}, \boldsymbol{\theta}_{\mathbf{n}}) = \frac{\exp\left[\sum_{k=1}^{K} b_{ik}\theta_{nk} + \sum_{p=1}^{P} a_{ip}\eta_p\right]}{1 + \exp\left[\sum_{k=1}^{K} b_{ik}\theta_{nk} + \sum_{p=1}^{P} a_{ip}\eta_p\right]}, \tag{1}$$

where

$X_{in} = 1$ if person n succeeds in solving item i $\quad (i = 1, \ldots, I; n = 1, \ldots, N)$,

$\mathbf{A} = (a_{ip})_{I \times P}$ is the matrix of integers,

$\mathbf{B} = (b_{ik})_{I \times K}$ is the matrix of positive integers,

$\boldsymbol{\eta} = (\eta_p)_{1 \times P}$ is the vector of basic parameters for the item stimulus features, and

$\boldsymbol{\theta}_{\mathbf{n}} = (\theta_{nk})_{1 \times K}$ is the ability vector of person n.

For a cognitive theory to be incorporated into the mathematical model, matrices **A** and **B** must be specified and can be understood as follows:

1. **A** is the $I \times P$ design matrix, with p denoting item features of a first type, $p = 1, \ldots, P$. The rows of **A** specify for each item the dependence of the difficulty on more basic parameters η_p, each corresponding to an item feature of the first type. Features of this type do contribute to the difficulty of the item. More in specific, the design matrix **A** defines the decomposition of the item parameters β_i into the P additive contributions of the basic parameters: η_p

$$\beta_i = \sum_{p=1}^{P} a_{ip}\eta_p.$$

2. **B** is the $I \times K$ score matrix, with K defining item features of a second type, $k = 1, \ldots, K$. Each feature of this second type is assumed to elicit a different type of individual differences playing a role in solving the item. The score weights b_{ik} specify for each item the dependence of a correct answer on source k of individual differences, called ability dimension θ_k, each of the θ_k corresponding to an item feature of the second type. Features of this type each correspond to a person ability. More specifically, the score matrix **B** defines the decomposition of the person's n contribution ξ_{in} to the probability of solving correctly item i into the K additive contributions of the person's abilities θ_{nk}, $\xi_{in} = \sum_{k=1}^{K} b_{ik}\theta_{nk}$.

A different assumption underlies both decompositions, however. By decomposing the item parameter β_i, one assumes that the effects of the item stimulus features (collected in **A**) on an item's difficulty are constant across persons. In contrast, by decomposing the persons' side, one assumes that the item stimulus features (collected in **B**) tap different abilities (assumption of multidimensionality), and hence that the features can have a different effect across persons on the probability of success. In fact, the design matrix defines a decomposition of item main effects, whereas the score matrix defines a decomposition of person main effects (if **B** contains a column of identical integers) and of person × item interaction effects (if **B** contains at least one column with nonidentical integers).

In principle, one can combine both approaches, as illustrated in Equation 1, the basic formula for the MRCMLM for the case of dichotomous responses. Often, however, one either models the cognitive theory only through a decomposition of the item parameters, as exemplified in the LLTM (Fischer, 1973; see Equation 2),

Downloaded from http://apm.sagepub.com at Lmog Ku Leuven Campusbiblio on April 11, 2008
© 2002 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

$$P\left(X_{in}=1 \mid \mathbf{A}, \boldsymbol{\eta}, \theta_n\right)=\frac{\exp\left[\theta_n+\sum_{p=1}^{P} a_{ip}\eta_p\right]}{1+\exp\left[\theta_n+\sum_{p=1}^{P} a_{ip}\eta_p\right]}, \tag{2}$$

or only through a decomposition of the person's side, as exemplified in the multidimensional polytomous latent trait model (MPLTM, Kelderman & Rijkes, 1994; see Equation 3 for the MPLTM for dichotomous items):

$$P\left(X_{in}=1 \mid \mathbf{B}, \beta_i, \boldsymbol{\theta}_{\mathbf{n}}\right)=\frac{\exp\left[\sum_{k=1}^{K} b_{ik}\theta_{nk}+\beta_i\right]}{1+\exp\left[\sum_{k=1}^{K} b_{ik}\theta_{nk}+\beta_i\right]}. \tag{3}$$

Kelderman and Rijkes (1994) express the numerator of (3) also as $\exp\left[\sum_{k=1}^{K} b_{ik}\left(\theta_{nk}+\delta_{ik}\right)\right]$, where $\sum_{k=1}^{K} b_{ik}\delta_{ik}=\beta_i$. δ_{ik} is the item parameter of item i with respect to ability θ_k. Hence, in the alternative expression, the item difficulty seems to be decomposed as well. However, the parameters estimated by the model are the β parameters. The individual δ_{ik} for the separate dimensions are in general not identifiable from those estimates (Kelderman & Rijkes, 1994, p. 162). If the b_{ik} are restricted to equal either zero or 1, the alternative formulation of the MPLTM is identical with the multicomponent model of Stegelmann (1983), who also recognizes the problem to estimate the individual δ parameters (though he refers to computational difficulties rather than to an identifiability problem, p. 261). He solves the problem by restricting the individual δ parameters to be equal across items for each dimension. The restricted multicomponent model is a special case of the model the authors propose, as they will discuss in the next section.

Both the LLTM and the MPLTM are special cases of the MRCMLM. For the LLTM, **B** is a one-column matrix of ones, and **A** specifies the cognitive theory. If **A** is the identity matrix, the Rasch model is obtained. For the MPLTM, **B** specifies the cognitive theory and **A** is the identity matrix (of rank *I*).

The elegance of the LLTM lies in its parsimony. However, the assumption that the effects of the item stimulus features are equal for all persons might be too stringent for some situations. As an example of such a situation, consider an item having both a verbal and a numerical aspect. One can expect that varying these aspects will differentially affect the difficulty of the item for persons with low versus high positions on the numerical or verbal ability. This person-dependent effect of an item stimulus feature on the probability of success cannot be captured by the LLTM.

The authors present a generalization of the LLTM that relaxes the assumption of equal feature weights for all persons. This random weights LLTM (RWLLTM) turns out to be a special case of the MRCMLM as well, in which each random weight corresponds with a separate dimension.

Downloaded from http://apm.sagepub.com at Lmog Ku Leuven Campusbiblio on April 11, 2008
© 2002 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

The Random Weights Linear Logistic Test Model

The Model

Taking a multilevel perspective on IRT modeling (Adams, Wilson, & Wu, 1997), one can consider the items of a test as being Level 1 units that are clustered within participants, the Level 2 units. The LLTM then can be seen as a multilevel logistic regression model (Hedeker & Gibbons, 1994; Zeger & Karim, 1991), as is more readily seen by expressing the LLTM in terms of the logit of the probability of success:

$$logit\,[P\,(X_{in} = 1 \mid \mathbf{A}, \boldsymbol{\eta}, \theta_n)] = \theta_n + \sum_{p=1}^{P} a_{ip}\eta_p. \tag{4}$$

In a multilevel perspective on the LLTM, the ability must be conceived as a random intercept (varying across the persons or Level 2 units), whereas the item stimulus features are modeled as fixed effects; their weights (slopes) are constant for all persons. In multilevel-models, it is common to assume a normal distribution for the random effects, but other distributions can be assumed as well (e.g., a nonparametric step-distribution, see the section "Estimation and Identifiability Issues").

Given a particular item, the (logistic) regression equation for two participants is the same, except for the intercept (ability). Hence, an assumption of the LLTM is that participants only differ with respect to their ability. In other terms: the LLTM allows for main effects of persons and item characteristics (subtasks), but not for interaction effects.

A straightforward generalization of the LLTM is to accept interactions between persons and (some) item characteristics so that the weights of these item characteristics can differ across persons. This is accomplished by considering (some of) the weights of the item characteristics as random effects. This model, the random weights linear logistic text model (RWLLTM), reads as

$$logit\,[P\,(X_{in} = 1 \mid \mathbf{A}, \boldsymbol{\eta}, \boldsymbol{\lambda}_\mathbf{n}, \theta_n)] = \theta_n + \sum_{p=1}^{P_1} a_{ip}\eta_p + \sum_{p=P_1+1}^{P} a_{ip}\lambda_{np}, \tag{5}$$

where

P_1 is the number of fixed effects,
$P_2 = P - P_1$ is the number of random effects,
$X_{in} = 1$ if person n succeeds in solving item i $\quad (i = 1, \ldots, I; n = 1, \ldots, N)$,
$\mathbf{A} = (a_{ip})_{I \times P}$ is the design matrix,
$\boldsymbol{\eta} = (\eta_p)_{1 \times P_1}$ is the vector of basic parameters of the fixed effects (equal for all persons),
$\boldsymbol{\lambda}_\mathbf{n} = (\lambda_{np})_{1 \times P_2}$ is the vector of basic parameters of the random effects (specific for person n), and
θ_n is the ability of person n (random intercept).

Alternatively, one can define the features 1 to P_1 to be features of the first type, and the features $P_1 + 1$ to P and the overall feature referring to all items (the intercept) as features of the second type, so that Equation 5 can be rewritten as Equation 1. More in specific, the design matrix ($\mathbf{A}$ in Equation 1) contains the P_1 columns corresponding with the fixed effects, and the score matrix ($\mathbf{B}$ in Equation 1) contains the $P_2 + 1$ columns (1 for θ_n) corresponding with the random effects; thus, θ_n in Equation 5 corresponds with θ_{n1} in Equation 1 (with $b_{i1} = 1$ for all i), the λ_{np} of Equation 5 correspond with the θ_{nk} in Equation 1 (with $k = p - P_1 + 1$), and the a_{ip} in Equation 5 correspond with the a_{ip} in Equation 1 for $p = 1, \ldots, P_1$, and with the b_{ik} for $p = P_1 + 1, \ldots, P$.

Hence, the RWLLTM simply is a MRCMLM, offering a clear understanding of what it means when an item feature is of the first or second type. The difference between the two approaches in

Downloaded from http://apm.sagepub.com at Lmog Ku Leuven Campusbiblio on April 11, 2008
© 2002 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

modeling a cognitive theory, considering item features as sources of item difficulty or as sources of individual differences, is simply the difference between the weights of the item features being person-independent and person-dependent, respectively.

The authors prefer to consider the RWLLTM as a separate model. An alternative would be to consider the RWLLTM as a way to interpret the score matrix of the MRCMLM. However, to stress the random weight LLTM interpretation of the score matrix, the authors prefer to label the RWLLTM as a separate model. The interpretation they propose is not evident. The MPLTM interpretation is more common: For example, when a cognitive test consists of different subtests, the score matrix is used to represent subtest membership of the items. More in general, the score matrix is thought to indicate which abilities play a role and is not considered to indicate which item features have a random effect to item difficulty. Therefore, the authors prefer to treat the RWLLTM as a separate submodel of the MRCMLM, as for example the common LLTM or the Rasch model.

Different interesting cases of the RWLLTM can be distinguished on the basis of the number of random effects:

1. $\mathbf{A} = \mathbf{0}$: only random effects. This is the case where all features contribute to individual differences. This is the restricted version of Stegelmann's (1983) multicomponent model.
2. $\mathbf{B} = \mathbf{0}$: only fixed effects. This case is the opposite of the previous one, with no individual differences but only item main effects.
3. $\mathbf{A} \neq \mathbf{0}$ and $\mathbf{B} \neq \mathbf{0}$: mixed effects. In this third case, some effects are random and some are fixed. The common LLTM is an example, with $\mathbf{B} = \mathbf{1}$, but other variants may be more realistic.

Estimation and Identification Issues

The results obtained for the MRCMLM can be applied directly to the RWLLTM, the latter being a particular specification of the former. The authors discuss shortly the marginal maximum likelihood (MML) estimation method. The latter is a natural approach from a multilevel perspective, in which (some of) the weights are considered to be random variables. For a more detailed account, the authors refer to Adams, Wilson, and Wang (1997). For the RWLLTM, the marginal likelihood that is maximized is

$$L(\boldsymbol{\eta}, \boldsymbol{\xi} \mid \mathbf{X}) = \prod_{n=1}^{N} \int_{\lambda} \prod_{i=1}^{I} \frac{\exp\left[x_{in}\left(\sum_{p=1}^{P_1} a_{ip}\eta_p + \sum_{p=P_1+1}^{P+1} b_{ip}\lambda_p\right)\right]}{1 + \exp\left(\sum_{p=1}^{P_1} a_{ip}\eta_p + \sum_{p=P_1+1}^{P+1} b_{ip}\lambda_p\right)} dG(\boldsymbol{\lambda} \mid \boldsymbol{\xi}), \tag{6}$$

where

$\mathbf{X} = (x_{in})_{I \times N}$ is the observed data matrix and

$G(\boldsymbol{\lambda} \mid \boldsymbol{\xi})$ is the the joint multivariate distribution function of the vector of random parameters characterized by the parameters $\boldsymbol{\xi}$.

Hence, the random weights are treated as nuisance parameters that are integrated out. A multivariate normal distribution for the random effects can be assumed, $\boldsymbol{\xi} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, but other distributions can be considered as well. Using a step distribution, with the "parameters" $\boldsymbol{\xi}$ being a prespecified set of nodes, the density at each of the nodes can be estimated from the data, so that an arbitrary continuous distribution can be approximated.

The maxima of the likelihood can be found by using Bock and Aitkin's (1981) formulation of the EM algorithm (Dempster, Laird, & Rubin, 1977). The algorithm requires an integration

Downloaded from http://apm.sagepub.com at Lmog Ku Leuven Campusbiblio on April 11, 2008
© 2002 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

over the marginal posterior distribution of the random weights, given the response patterns. In the step-distribution case, the integral becomes a summation. In the normal case, the integration can be approximated by Gaussian (multidimensional) quadrature (Bock & Aitkin, 1981; Gibbons & Bock, 1987). However, as the number of random effects increases, the number of terms in the summation approximating the integral increases exponentially (if there are Q nodes for each of the P_2 dimensions, the summation is over Q^{P_2} terms). Fortunately, Bock, Gibbons, and Muraki (1988) found that, with increasing dimensionality, the number of nodes per dimension can be reduced without impairing the accuracy; for a five-dimensional solution, only three nodes per dimension was sufficient.

Adams et al. (1997) and Volodin and Adams (1995) formulated necessary and sufficient conditions with respect to the design matrix **A** and the score matrix **B** for the MRCMLM to be identified. For the particular case of the RWLLTM, these identifications conditions are as follows:

1. $P + 1 \leq I$ (that is, the total number of effects should not exceed the number of items).
2. $rank(\mathbf{A}) = P_1$; $rank(\mathbf{B}) = P_2 + 1$; $rank([\mathbf{A}\ \ \mathbf{B}]) = P + 1$ (that is, the design and score matrix should be of full-column rank, as well as the concatenation of both matrices).
3. $rank([\mathbf{A}\ \ \mathbf{B}]) = P + 1 \leq I$ (a consequence of 1 and 2).

Furthermore, the covariance matrix of the distribution of the random effects has to be symmetric positive definite for the model to be identified. This condition is not fulfilled if the random effects of a particular model are (almost) linear dependent, or if the variance of a random effect equals (almost) zero (hence, if it is in fact a fixed effect). A possible solution for either of the two cases is to consider a model with one random effect variable less.

Once a model is estimated, its fit can be evaluated by looking at the approximate standard normal fit statistic (Wu, 1997; Wu, Adams, & Wilson, 1998) that can be computed for the basic parameters of the predictors of the design matrix. The statistic is based on weighted standardized residuals and is a generalization of the Wright and Masters (1982) fit statistic (see also Hoskens & de Boeck, 2001, for the use of this statistic).

Different models that are nested can be contrasted using the likelihood ratio test statistic, whereas different models that are not nested can be compared using information criteria such as Akaike's (1974) information criterion (AIC) or the Bayesian information criterion (BIC; Schwarz, 1978). However, if one wants to test a model with P random effects against a model with $P + 1$ random effects, a likelihood ratio test is not appropriate, even though the former is nested within the latter. This is because the nested model is on the boundary of the parameter space of the alternative model (by setting one variance to zero). The asymptotic chi-squared null distribution for the likelihood ratio test statistic is not necessarily valid any longer. In that case, the information criteria may be indicative as well.

Relation With Other IRT Models

The RWLLTM is related to some existing IRT models that also incorporate the idea of multilevel modeling. The RWLLTM differs from the latter with respect to which aspect of the model is considered to be of a random nature. In the following, the authors sketch the outlines of two related IRT models, and the differences with their approach will become apparent. The two models differ with respect to the randomness of the coefficients. In the first, the weights are random over persons, and in the second, they are random over items within an item group.

First, Reise (2000) uses multilevel logistic regression to evaluate the fit of IRT models. His approach consists of two steps. First, item and person parameters are estimated for a particular IRT

Downloaded from http://apm.sagepub.com at Lmog Ku Leuven Campusbiblio on April 11, 2008
© 2002 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

model. For example, for the 2-parameter logistic model (in terms of the logit),

$$logit\,[P(X_{in} = 1 \mid \alpha_i, \beta_i, \theta_n)] = \alpha_i(\theta_n - \beta_i), \tag{7}$$

the parameters are the item difficulty parameters β_i, the item discrimination parameters α_i, and the person ability parameters θ_n. Second, the estimates obtained in the first step are used as predictors for the following multilevel logistic regression model:

$$logit\,[P(X_{in} = 1 \mid \beta_i)] = b_{0n} + b_{1n}\beta_i. \tag{8}$$

The first-level random intercept b_{0n} and random weight b_{1n} in turn are modeled on a second level as

$$b_{0n} = \gamma_{00} + \gamma_{01}\theta_n + \delta_{0n}, \tag{9a}$$

$$b_{1n} = \gamma_{10} + \delta_{1n}. \tag{9b}$$

The δ-parameters are normally distributed error terms; if dropped, the intercept b_{0n} and weight b_{1n} no longer are treated as random effects, as their variance becomes zero.

If the estimated variance of δ_{0n} is substantial, this might indicate a violation of the local independence assumption (e.g., due to multidimensionality) of the 2-parameter logistic model, because according to the latter, the intercept b_{0n} should vary as a function of ability θ_n only. If the estimated variance of δ_{1n} is substantial, the slope of item difficulty (b_{1n}) differs over persons, meaning that people differ in the degree to which the item difficulties affect the probability of success, or in other words, that the items are not discriminating to the same degree for all persons. According to the 2-parameter logistic model, each item should discriminate to the same degree for all persons, and hence the estimated variance of δ_{1n} offers another possibility to evaluate the fit of the model. The main differences between Reise's (2000) approach and the RWLLTM are the following: First, Reise follows a two-step approach, with the item and person parameters estimates obtained in the first step functioning as predictors in a second step; he applies a multilevel model to the estimates of a 2-parameter logistic model. The RWLLTM is a one-step approach; all parameters are estimated at once. Second, Reise uses multilevel models to evaluate IRT models, whereas the RWLLTM is a multilevel logistic regression model.

Second, the starting point of the random-effects LLTM of Janssen and De Boeck (2002) is the observation that the LLTM, when tested, often does not fit the data even if the predictors of the design matrix explain most of the variation in the item difficulties. Janssen and De Boeck relax the LLTM by adding an item-specific error term to the decomposition of the item difficulty in terms of the basic parameters:

$$\beta_i = \sum_{p=1}^{P} a_{ip}\eta_p + \varepsilon_i.$$

The error term is assumed to be normally distributed with a mean of zero. Hence, the item difficulties are considered to be random effects. By adding the error term, it is no longer required that items with the same combination of values on the predictors (items of the same "item group") are of equal difficulty. Differences in difficulty between items of the same family may occur due to item-specific characteristics, such as wording or content of the items. In contrast with the RWLLTM in which the effects are random across persons, the effects in the random-effects LLTM are random across items belonging to the same item group.

The approach of considering items as being clustered within an item group is also followed in the context of the normal ogive model (the probit version of the 2-parameter logistic model) by Janssen, Tuerlinckx, Meulders, and De Boeck (2000).

Downloaded from http://apm.sagepub.com at Lmog Ku Leuven Campusbiblio on April 11, 2008
© 2002 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

Illustration: Deductive Reasoning

The example is based on a study of Rijmen and de Boeck (2001) in which propositional reasoning problems were constructed as a combination of more elementary inferences (item stimulus features). IRT modeling for deductive reasoning has shown to be successful in an earlier study by Heinrich (1975). Heinrich used syllogisms (e.g., "All men are mortal"; "Socrates is a man"; hence, "Socrates is mortal") to investigate the effects of familiarity (e.g., "all mammals are animals" vs. "all screeps are heeps") and veracity of the premises (e.g., "all monkeys are mammals" vs. "all monkeys are birds"). She found that the effect of familiarity could be measured independent of the logical structure, whereas the effect of veracity could not. In their study, the authors are primarily interested in how the difficulty of complex reasoning problems can be accounted for by its basic logical structure. Therefore, they use propositional reasoning problems (e.g., "If Socrates is a man, then Socrates is mortal"; "Socrates is a man"; hence, "Socrates is mortal"), with a compound structure, as will be explained in the next section.

Method

Participants

Two hundred fourteen high school students between 16 and 19 years of age participated in the experiment.

Material

Each participant received a booklet, with 2 pages of instructions and 10 pages with problems. On each page, 3 problems were presented, for a total of 30 problems. Under each problem, 4 response alternatives were presented:

- necessarily true
- necessarily not true
- undecidable
- I don't know

It was explained in the instructions that participants should choose "necessarily true" if they believed that, given true premises, the conclusion must be true; "necessarily false" if they believed that the conclusion must be false; "undecidable" if they believed that, given the premises, the conclusion could be true, but also could be false, hence, that not enough information was given; and "I don't know" if they could not work it out. It was explained to the participants that "undecidable" is a genuine conclusion that could be correct for some of the problems, whereas "I don't know" means one cannot choose among the other three response alternatives or one has given up solving the problem. The "I don't know" was offered as an alternative response in order to avoid guessing behavior. Four random sequences were generated for the problems, equally divided over the participants.

Design

Each participant had to evaluate the conclusion of 30 complex problems. The correct answer on 24 of them was independent of the interpretation of "IF/THEN" (conditional vs. biconditional). These were the items of primary interest. For the other six items, the correct answer was dependent on the interpretation of "IF/THEN." These were filler items and are not considered further. The 24 problems of interest were constructed as a combination of elementary inferences: conjunction (p, q $\therefore$ p AND q), modus ponens (IF p, THEN q; p $\therefore$ q), disjunctive modus ponens (IF p OR q, THEN r; p $\therefore$ r), disjunctive syllogism (p OR q; NOT p $\therefore$ q), solving a contradiction (p, NOT p $\therefore$ incompatible), and modus tollens (IF p, THEN q; NOT q $\therefore$ NOT p).

Downloaded from http://apm.sagepub.com at Lmog Ku Leuven Campusbiblio on April 11, 2008
© 2002 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

The 24 problems were constructed by combining premises as illustrated in Table 1, resulting in 6 problem types. A letter represents an elementary proposition, possibly containing a negation. It was never necessary to perform a double negation (e.g., for Problem Type 3 of Table 1, "q" never represented a proposition with a negation).

Table 1
The Six Problem Types

	Modus ponens	Modus tollens
Modus ponens + conjunction	1. IF p, THEN q p IF r, THEN s r ——————— q AND s	2. IF p, THEN q p IF r, THEN s NOT s ——————— q AND NOT r
Disjunctive syllogism	3. p OR q IF r, THEN NOT q r ——————— p	4. p OR q IF q, THEN r NOT r ——————— p
Disjunctive modus ponens	5. IF p OR q, THEN r IF s, THEN q s ——————— r	6. IF p OR NOT q, THEN r IF q, THEN s NOT s ——————— r

Modus ponens versus modus tollens was the first factor in the design. For 12 reasoning problems, one inference was a modus ponens (left column). For the other 12, the corresponding inference was a modus tollens (right column).

The second factor of the design refers to the inferences that are combined with the first modus ponens or modus tollens inference and had three levels: modus ponens + conjunction, disjunctive syllogism, and disjunctive modus ponens. This factor was orthogonal with the first and makes a distinction between three sets of eight problems. For the first set of eight reasoning problems, problems of level modus ponens + conjunction (first row of Table 1), the second inference was a modus ponens. This modus ponens inference could be made independently from the modus ponens or modus tollens inference already discussed. The results of both conditional inferences subsequently had to be combined with a conjunction, which was then the third inference. For the second set of eight problems, problems of level disjunctive syllogism (second row of Table 1), the second (and last) inference to be made was a disjunctive syllogism. The categorical proposition of the disjunctive syllogism was the result of the modus ponens or modus tollens inference from the first step. The third set of eight problems, problems of level disjunctive modus ponens (third row of Table 1), was similar to the second set, except for the second inference now being a disjunctive modus ponens.

A third factor, orthogonal to the first two, was the truth-value of the conclusion to be evaluated: true or false. For 12 problems, the conclusion to be evaluated was true; for the other 12, the conclusion to be evaluated was false. For the latter, the conclusion to be evaluated contradicted the conclusion that followed from the premises. Solving this contradiction was an additional inference for these problems. To avoid the presence of negations being a cue for the correct answer, some propositions of the premises contained a negation, though none of the problems required a double negation inference.

Downloaded from http://apm.sagepub.com at Lmog Ku Leuven Campusbiblio on April 11, 2008
© 2002 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

The fourth factor, manipulated orthogonally, was the content of the problems. Half of the problems were about people in cities or countries, for example, "John is in Paris." The other half were about a functioning factory, for example, "The green light flashes." The orthogonal manipulation of these factors resulted in $2 \times 3 \times 2 \times 2 = 24$ cells in the design, each corresponding to exactly one problem.

A fifth factor, not orthogonally manipulated because this would render the experiment too long, was the order of presentation of the premises: grouped versus ungrouped. For each type of content, the premises of a given inference were presented as adjacent premises in half of the problems (as presented in Table 1), and separated by another premise in the remaining half of the problems (e.g., for Type 1: "p" became the first premise and "IF p, THEN q" the third, with the premise "IF r, THEN s" of the other modus ponens in between). For the content of a factory, these problems were of Types 2, 3, and 6, and for the content of people in cities or countries, these problems were of Types 1, 4, and 5.

Procedure

The experiment was conducted in groups of more or less 20 participants. At the beginning of the experiment, the instructions were briefly explained. Participants were asked not to guess but to choose the response alternative "I don't know" in cases where they could not make one of the three evaluations: necessarily true, necessarily not true, or undecidable. It was stressed that they were not allowed to write anything down while solving the problems. Participants had to complete the task within 50 minutes, and they all succeeded in doing so.

Results and Discussion

The difficulties of the 24 problems, expressed in proportion correct answers, ranged from .44 to .98, with a mean of .72.

A preliminary regression analysis was performed on the logit of the proportion correct answers of the 24 problems. The resulting estimates of the regression weights can serve as a basis for comparison for the estimates of the generalized RWLLTM. The three factors of the design referring to elementary inferences were coded into four binary variables as follows:

P1: 1/0 if first inference modus ponens/modus tollens
P2: 1/0 if further inference modus ponens + conjunction/otherwise
P3: 1/0 if further inference disjunctive syllogism/otherwise
P4: 1/0 if conclusion false/true

Problems with a value of zero on both P2 and P3 were problems with a disjunctive modus ponens as second inference.

The four predictors P1 to P4 explained 74% of the variance of the logit transformation of the proportion correct answers, $F(4,19) = 13.82$; $p < .001$. The (standardized) regression weights for P1 to P4, in descending order of absolute value were 1.09 (.51) for P1, $t(23) = 4.38$, $p < .001$; –.86 (–.38) for P3, $t(23) = -2.81$, $p < .05$; .80 (.35) for P2, $t(23) = 2.62$, $p < .05$; and –.64 (–.30) for P4, $t(23) = -2.58$, $p < .05$. Positive coefficients indicate that problems with a coding of 1 were easier.

To test the effect of the other factors of the design (not referring to elementary inferences: content and presentation order of the premises), these factors were also coded into binary variables:

P5: 1/0 if content of functioning factory/content of people in cities or countries
P6: 1/0 if premises of the same inference not presented together/presented together

Downloaded from http://apm.sagepub.com at Lmog Ku Leuven Campusbiblio on April 11, 2008
© 2002 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

When either P6 or P5 are added to the predictor set, these factors were not significant with an alpha level of .05, $t(23) = .74$, $p = .47$; and $t(23) = -1.70$, $p = .11$, respectively). Therefore, P5 and P6 are omitted in the remainder of the analyses.

For all the RWLLTM models the authors tested, they assumed a multivariate normal distribution for the random effects. The models were estimated with the marginal maximum likelihood method, as described earlier on in the article. The estimation was done with the CONQUEST software (Wu et al., 1998). For the models with one and two random effects, 20 nodes were used for the Gaussian quadrature. To reduce computing time, the number of nodes was reduced to 10 and 8, respectively, for the models with three and four random effects, and for the models with five random effects.

For technical reasons, the authors reparameterized λ_{np} of Equation 5 as $\lambda_{np} = \lambda^*_{np} + \overline{\lambda_{np}}$, with $\lambda^*_{np} = \lambda_{np} - \overline{\lambda_{np}}$. Hence, the random coefficient λ_{np} was defined as into the sum of a fixed coefficient $\overline{\lambda_{np}}$ and a random coefficient λ^*_{np} with mean equal to zero. This means that the random part of the model now stands for the deviations from the mean difficulty of the item stimulus features. Hence, the mean of the random effects are zero by definition. The advantages of this technical operation were twofold. First, the authors no longer had to specify a separate design matrix for each model, but instead could use the same design matrix for all RWLLTM models, the estimated coefficients being either the coefficients of the "true" fixed effects (the η's) or the estimates of the means of the random effects (the $\overline{\lambda_{np}}$'s). Second, the fit of the models could be evaluated better because CONQUEST provides fit statistics for the basic parameters of the predictors of the design matrix (fixed effects; c.q. the approximate standard normal fit statistics discussed earlier), but not for the estimated means of the latent distribution (means of the random effects). By putting the means of the random effects into the fixed part of the model, it nevertheless was possible to obtain fit statistics for the estimated means of the latent distribution.

For all models the authors tested, the same design matrix (**A** in Equation 1) was used, consisting of columns that correspond to P1, P2, P3, P4, and a column of ones (for the intercept). As explained, this can be done irrespective of the weights considered being fixed or random. The models differ as to the score matrix (**B** in Equation 1) the authors used. In all cases, a 1-vector was used corresponding to the random intercept. Furthermore, for each effect that was considered random, the corresponding binary coded variable (P1, P2, P3, P4) was added as a column in the score matrix.

In total, 16 different models of five types were estimated: (a) one model with the intercept as the only random effect (the LLTM); (b) four models with the intercept and the weight of one of the four binary predictors as random effects (hence, the four models differed with respect to which was the second random effect besides the intercept); (c) six models with the intercept and the weights for two of the four binary predictors as random effects (hence, the six models differed with respect to which were the second and third random effect besides the intercept); (d) four models with the intercept and the weights for three of the four binary predictors as random effects; (e) and finally one model with all effects random. Several models were not estimable due to the estimated covariance matrix of the population model becoming singular during the estimation process, indicating that there was no variation in the weight of a predictor across participants, or that the random weights were not linearly independent. Considering all 16 models is a blind (automatic) selection procedure, to be contrasted with a hypothesis-driven selection procedure described next.

Not all models can be considered to be plausible on the basis of the relevant literature. Indeed, in the deductive reasoning literature, differences are reported mainly with respect to the interpretation of disjunctions and conditionals (for a review, see Evans, Newstead, & Byrne, 1993). A disjunction can be understood exclusively (either p or q) or inclusively (p or q, or both). A conditional can be understood as a pure conditional (p implies q, but not vice versa) or as a biconditional (p implies q, and vice versa). The particular interpretation one adopts influences the difficulty of the reasoning

Downloaded from http://apm.sagepub.com at Lmog Ku Leuven Campusbiblio on April 11, 2008
© 2002 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

operations involved, respectively disjunctive syllogism and modus tollens (Braine & O'Brien, 1991; Johnson-Laird, Byrne, & Schaeken, 1992). Hence, if a model with random weights is selected, the authors expect it to be a model with P1, P3, or both as random effects in addition to the random intercept. They do not expect random effects for P2 and P4. In this respect, it is no surprise that some of the models were not estimable due to a singular covariance matrix for the random effects, which should be the case if there is no variation in the weights of one of the item features.

The authors did opt for the blind (automatic) selection procedure because they were interested in comparing the selected model with all other models, and especially with those having the same number of random effects. In combination with the expectations derived from the literature, the blind (automatic) procedure is a way to make sure one has found the best model. If the blind procedure finds what is within the range of expectation, then one can trust the result all the more.

For each number of random effects, the best fitting model (in terms of –2 × log likelihood) was selected. These models are given in Table 2, together with their estimated covariance matrix. For each of these best fitting models, the authors computed the value of the BIC (see Table 2). In terms of the BIC, the model with as random effects the intercept together with the weight of P3 (disjunctive syllogism vs. modus ponens and conjunction or disjunctive modus ponens) is the model to be selected. The –2 × log likelihood of this model is only marginally higher than that for the model with three random effects (intercept, P2 and P3). The values are, respectively, 5212.94 and 5212.90. The **A** and **B** matrices of the selected model are partially given in Table 3 for the first eight problems.

Table 2
Fit Indices and Variance-Covariance Matrices for the Best-Fitting Random Weights Linear Logistic Model

Random Effect	−2 × Log Likelihood	BIC	Covariance Matrix				
			Intercept	P1	P2	P3	P4
Intercept	5284.54	5316.74	.65				
Intercept + P3[a]	5212.94	5255.87	.51			.02	
						1.24	
Intercept + P2	5212.90	5271.93	.67		−.23	−.04	
+ P3					.63	.06	
						1.20	
Intercept + P1	5180.60	5261.09	1.05	−.57	−.26	−.33	
to P3				.73	.07	.51	
					.68	.17	
						1.46	
Intercept + P1 to P4	Singular covariance matrix						

Note. BIC = Bayes Information Criterion
[a]Selected model.

The estimates of the fixed effects and the means of the random effects are comparable in sign and relative magnitude with the regression weights of the multiple regression (see Table 4), supporting the adequacy of the model. Again, positive estimates indicate that problems with a coding of 1 were easier. Hence, problems in which a modus ponens has to be made are easier than problems in which a modus tollens has to be made instead; problems in which a modus ponens and a conjunction have to be made are easier than problems with a disjunctive modus ponens or a disjunctive syllogism instead; problems in which a disjunctive syllogism has to be made are more difficult than problems with a disjunctive modus ponens or a modus ponens and a conjunction instead; and finally, problems with a

Downloaded from http://apm.sagepub.com at Lmog Ku Leuven Campusbiblio on April 11, 2008
© 2002 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

Table 3
Design Matrix **A** and Score Matrix **B** of the
Selected Model for the First Eight Problems

	Design Matrix **A**					Score Matrix **B**	
Problem	Intercept	P1	P2	P3	P4	Intercept	P3
1	1	1	1	0	1	1	0
2	1	1	1	0	1	1	0
3	1	0	0	1	0	1	1
4	1	0	0	0	0	1	0
5	1	0	0	0	1	1	0
6	1	1	0	1	0	1	1
7	1	0	0	1	1	1	1
8	1	0	1	0	0	1	0

false conclusion are more difficult than problems with a true conclusion. The estimates, which refer to elementary inferences that are part of more complex problems, are consistent with the results of experiments in which the elementary inferences are studied separately (Johnson-Laird et al., 1992; for a review, see Evans et al., 1993). Using a multinomial model approach (Batchelder & Riefer, 1990; Hu & Batchelder, 1994), Klauer and Oberauer (1995) also found that modus ponens is more readily endorsed than either disjunctive syllogism or modus tollens.

The estimated variances were .51 and 1.24, respectively, for the intercept and the weight of P3. The estimated covariance between the intercept and the weight of P3 is .02. Apparently, there is a large variation in the difficulty of a disjunctive syllogism across participants, which is unrelated with the variation in the intercept. The large variance of the weight of P3 is also found in the other best fitting models (see Table 2). Hence, a disjunctive syllogism renders a problem rather difficult (by the negative estimate for the mean effect of P3), but more so for some participants than for others (by the large variance for the person-specific deviations from the mean effect of P3). The common LLTM is incapable of grasping the interaction between participants and P3. It provides a considerably worse fit than the model with P3 as a random effect (see Table 2: $-2 \times$ log likelihood $= 5284.54$ and 5212.94, respectively).

The approximate standard normal fit statistics indicated a good fit except for a moderate misfit for the estimate of the weight for P1 (first inference modus ponens/modus tollens; see Table 4), with a value of 2.83, $p = .005$. This may indicate that there exists also some variation in the weight of P1 across persons. Unfortunately, the model with the intercept random, as well as both the weights of P1 and P3, was not estimable, due to the covariance matrix becoming singular during the estimation process. This probably indicates that a dimensionality of three (for the intercept, and for the weights of P1 and P3) is too high and that, therefore, the extension to a random weight for P1 is not a good idea.

Concluding Remarks

There are three ways to look upon the RWLLTM the authors proposed. First, the RWLLTM offers a combination of two approaches to model data starting from a cognitive theory, just like the MRCMLM of which the RWLLTM is a special case. For both approaches, the underlying assumption is that a cognitive task can be characterized by a set of item stimulus features. Different approaches differ with respect to the way these stimulus features are incorporated in the mathematical model. In the LLTM, the item contribution is decomposed into a weighted sum of the stimulus features, and the ability is not further analyzed. In the MPLTM, on the other hand, the contribution

Downloaded from http://apm.sagepub.com at Lmog Ku Leuven Campusbiblio on April 11, 2008
© 2002 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

Table 4
Estimates of the Fixed Effects and the Corresponding Multiple Regression Weights

Parameter	Weight in Multiple Regression	Estimate in RWLLTM	Approximate Standard Normal Fit Statistic
Intercept	1.003	.970[a]	−1.31
P1	1.093	1.030	2.83*
P2	.802	.852	.89
P3	−.858	−.700[a]	−.17
P4	−.645	−.578	.41

Note. RWLLTM = Random weights linear logistic model.
[a]Mean of random effects.
*$p < .01$.

of the person is decomposed based on the stimulus features, and the item difficulty is not further analyzed. The authors' approach combines the characteristics of the two approaches: decomposing the item side as in the LLTM, and decomposing the person abilities, as in the MPLTM.

Second, the RWLLTM is an individual-differences generalization of the LLTM, with the weights of the item stimulus features as random coefficients. As such, the RWLLTM is an individual-differences extension of the LLTM.

Third, the RWLLTM can also be considered as a model for differential item functioning, in that it allows for interactions between persons and subtasks (and consequently between persons and items). However, the RWLLTM addresses differential item functioning depending on the individual, and not on the group the individual belongs to. Usually, one looks for differential item functioning across groups of persons.

The illustrative example showed that the RWLLTM was indeed in closer correspondence with the data than the common LLTM. The weights of one item feature turned out to be a major source of individual differences, even much larger than the individual differences with respect to the intercept referring to a more general ability. However, the common LLTM only takes the latter into account. By allowing one or a few weights to vary over persons, the RWLLTM offers a more flexible approach.

References

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

Adams, R. J., Wilson, M., & Wu, M. L. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics, 22,* 47-76.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716-723.

Batchelder, W. H., & Riefer, D. M. (1990). Multinomial models of source monitoring. *Psychological Review, 97*, 548-642.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika, 46*, 443-459.

Bock, R. D., Gibbons, R. D., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12*, 261-280.

Braine, M. D. S., & O'Brien, D. P. (1991). A theory of if: A lexical entry, reasoning program, and pragmatic principles. *Psychological Review, 98*, 182-203.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39*, 1-38.

Embretson, S. E. (1984). A general multicomponent latent trait model for response processes. *Psychometrika, 49*, 175-186.

Downloaded from http://apm.sagepub.com at Lmog Ku Leuven Campusbiblio on April 11, 2008
© 2002 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.

Embretson, S. E. (1985). *Test design: Developments in psychology and psychometrics.* New York: Academic Press.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*, 380-396.

Evans, J. St. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning: The psychology of deduction.* Hove, UK: Erlbaum.

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

Gibbons, R. D., & Bock, R. D. (1987). Trend in correlated proportions. *Psychometrika, 52*, 113-124.

Hedeker, D., & Gibbons, R. D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics, 50*, 933-944.

Heinrich, I. (1975). Changes in syllogistic reasoning influenced by semantic variation of propositions. *Psychologische Beiträge, 17*(4), 497-518.

Hoskens, H., & de Boeck, P. (2001). Multidimensional componential item response theory models for polytomous items. *Applied Psychological Measurement, 25*, 19-37.

Hu, X., & Batchelder, W. H. (1994). The statistical analysis of general processing tree models with the EM algorithm. *Psychometrika, 59*, 21-47.

Janssen, R., & De Boeck, P. (2000). *A random-effects version of the linear logistic test model.* Manuscript submitted for publication.

Janssen, R., Tuerlinckx, F., Meulders, M., & De Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics, 25*, 285-306.

Johnson-Laird, P. N., Byrne, R. M. J., & Schaeken, W. (1992). Propositional reasoning by model. *Psychological Review, 99*, 418-439.

Kelderman, H., & Rijkes, C. P. M. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika, 59*, 149-176.

Klauer, K. C., & Oberauer, K. (1995). Testing the mental model theory of propositional reasoning. *The Quarterly Journal of Experimental Psychology, 48A*, 671-687.

Linacre, J. M. (1994). *Many-facet Rasch measurement.* Chicago: MESA Press. (Original work published 1989)

Maris, E. (1995). Psychometric latent response models. *Psychometrika, 60*, 523-547.

Reise, S. P. (2000). Using multilevel logistic regression to evaluate person-fit in IRT models. *Multivariate Behavioral Research, 35*, 545-570.

Rijmen, F., & de Boeck, P. (2001). Propositional reasoning: The differential contribution of "rules" to the difficulty of complex reasoning problems. *Memory & Cognition, 29*, 165-175.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461-464.

Stegelmann, W. (1983). Expanding the Rasch Model to a general model having more than one dimension. *Psychometrika, 48*, 259-267.

Volodin, N. A., & Adams, R. J. (1995, April). *Identifying and estimating a D-dimensional item response model.* Paper presented at the International Objective Measurement Workshop, University of California, Berkeley.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement.* Chicago: MESA Press.

Wu, M. (1997). *The development and application of a fit test for marginal maximum likelihood estimation and generalized item response models.* Unpublished master's dissertation, University of Melbourne.

Wu, M. L., Adams, R. J., & Wilson, M. (1998). *Acer CONQUEST: Generalized Item Response Modelling Software manual*. Melbourne, Victoria: The Australian Council for Educational Research Ltd.

Zeger, S., & Karim, R. (1991). Generalized linear models with random effects. *Journal of the American Statistical Association, 86*, 79-86.

Acknowledgments

Frank Rijmen was supported by the Fund for Scientific Research Flanders (FWO). The authors would like to thank anonymous reviewers for their useful comments on an earlier draft.

Author's Address

Send requests for reprints or further information to Frank Rijmen, Afdeling Psychodiagnostiek, University of Leuven, Tiensestraat 102, 3000 Leuven, Belgium. E-mail: frank.rijmen@psy.kuleuven.ac.be.

Downloaded from http://apm.sagepub.com at Lmog Ku Leuven Campusbiblio on April 11, 2008
© 2002 SAGE Publications. All rights reserved. Not for commercial use or unauthorized distribution.