Applied Psychological Measurement

http://apm.sagepub.com

Locally Dependent Linear Logistic Test Model With Person Covariates

Edward H. Ip, Dirk J. M. Smits and Paul De Boeck

Applied Psychological Measurement 2009; 33; 555
DOI: 10.1177/0146621608326424

The online version of this article can be found at:
http://apm.sagepub.com/cgi/content/abstract/33/7/555

Published by:



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 http://apm.sagepub.com/cgi/content/refs/33/7/555

Downloaded from http://apm.sagepub.com at Katholieke Universiteit Leuven on December 16, 2009

Applied Psychological Measurement
Volume 33 Number 7
October 2009 555-569
© 2009 Sage Publications
10.1177/0146621608326424
http://apm.sagepub.com
hosted at
http://online.sagepub.com

Locally Dependent Linear Logistic Test Model With Person Covariates

Edward H. Ip
Wake Forest University School of Medicine
Dirk J. M. Smits
University College Brussels
Paul De Boeck
K. U. Leuven

The article proposes a family of item-response models that allow the separate and independent specification of three orthogonal components: item attribute, person covariate, and local item dependence. Special interest lies in extending the linear logistic test model, which is commonly used to measure item attributes, to tests with embedded item clusters. The problem of local item dependence arises in item clusters. Existing methods for handling such dependence, however, often fail to satisfy the property of invariant marginal interpretation of the item attribute parameters. Although such a property may not be necessary for applications that focus on predictive analysis, it is critical for linear logistic test models. To achieve the marginal property, we implement an iterative estimation method, which is illustrated using data collected from an inventory on verbal aggressiveness.

Keywords: *Index terms: Item response theory, item attribute, person covariate, local item dependence, item cluster, testlet, linear logistic test model*

One of the most important contributions of item-response theory (IRT) to the fields of psychology, education, and the behavioral sciences has been the delineation of item–person interaction. As observed by Glas (2005), following the pioneering work of Rasch (1960), Lord (1980), and others, generalization of the principles of IRT, especially in the study of individual differences, has evolved in three directions: (a) inclusion of item components in characterizing the item parameter, (b) introduction of person attributes into model differences between subgroups, and (c) inclusion of local item dependencies (LID) of item responses that arise in response contexts and processes that are beyond the explanation of the latent trait. This article describes a locally dependent linear logistic test model (LID-LLTM) that allows the separate and independent specification of the three components described by Glas: (a) item attributes, (b) person covariates, and (c) an LID that is not captured by person trait. Most important, the LID-LLTM adopts a marginal approach (described later) that preserves the interpretation of components (a) and (b) in the same way that they are interpreted in standard IRT models. Therefore, the proposed approach adds to the existing literature by providing a proper measurement framework that allows the incorporation of all three components while not altering the respective interpretations of the first two, which are typically the points of interest for practitioners and researchers.

Downloaded from http://apm.sagepub.com at Katholieke Universiteit Leuven on December 16, 2009

A prominent example of the first direction outlined by Glas is the LLTM proposed by Fischer (1973) and its various extensions (Embretson, 1984; Fischer, 1995; Janssen & De Boeck, 1997; Rijmen & De Boeck, 2002; Wilson, 1989). The LLTM allows the direct assessment of the impact of item attributes on the difficulty of an item. Item attributes, particularly in applications such as cognitive diagnosis, can provide more meaningful interpretations than can item-difficulty parameters. For the second direction, person attributes such as gender, age, and other measures of personal traits can be introduced into IRT as multilevel models (Mislevy & Sheehan, 1989; Zinderman, 1997).

The last direction, broadly termed LID, has been the object of growing interest, presumably in response to a demand for more flexible IRT models to handle complex processes and test designs. The LID arises when items tend to cluster within a test. For example, an item cluster comprises items that are related to a common reading passage in an educational test or items that are related to the same hypothetical situation in a psychological test. The information that is contained in an item within a cluster for the measurement of the intended construct needs to be discounted, and the extent to which it needs to be discounted depends on the level of the LID.

One of the most fruitful areas of investigation of LID is in the development of formal models for both modeling and explanation of LID. This includes random-effects (RE) models (Bradlow, Wainer, & Wang, 1999; for RE testlet response models, see Wainer, Bradlow, & Wang, 2007; for RE facet models, see Wang & Wilson, 2005; for multiscale RE models, see Scott & Ip, 2002); componential models (Hoskens & De Boeck, 1997); item-bundle models (Rosenbaum, 1988); marginal and hybrid models (Ip, 2000, 2002; Ip, Wang, De Boeck, & Meulders, 2004); bifactor models (Gibbons & Hedekker, 1992); log-linear models (Jannarone, 1986; Kelderman & Rijkes, 1994); and models with internal restrictions on item difficulties (MIRID; Butter, De Boeck, & Verhelst, 1998). The remarkably broad range of LID models for educational and psychological testing reflects a growing need for measurement technologies in which traditional assumptions such as local independence may not necessarily be valid for studying individual differences.

Background

The most important methodological challenge of constructing an encompassing model for (a)–(c), as one sees it, is to maintain, in the presence of LID, the marginal interpretation (which shall be explained later) of the item parameter within the LLTM component of the model. Mathematically, the issue is related to the orthogonality among the components. Many existing LID models do not maintain the orthogonality condition, and thus they would not be entirely suitable for the LLTM.

The LLTM extends the Rasch model by allowing the kernel of the response function (logit of probability of positive response given ability) to be decomposed into the sum of person trait and a compensatory term of some weighted linear function of item attributes. In other words, in the LLTM, the item difficulty parameter can be expressed by a design matrix of items by components, multiplied by the basic component parameters. Maintaining this structure is a nontrivial task if it is necessary to account for LID, and one of the

Downloaded from http://apm.sagepub.com at Katholieke Universiteit Leuven on December 16, 2009

Table 1
Joint Probabilities of Response Patterns of Three LID Items With Given Marginal ICCs

Response Pattern	100	110	101	111	000	010	001	011
Probability under local independence $P(.)$	0.0864	0.0864	0.1215	0.1215	0.1215	0.1215	0.1706	0.1706
Probability under LID $P^*(.)$	0.0247	0.0247	0.0734	0.2931	0.2931	0.0732	0.1088	0.1088

Note: Item parameter values $b = (0.2, 0, -0, 2)$. Pairwise conditional odds ratios between item pairs (1, 2), (1, 3), and (2, 3) are, respectively, 4, 8, and 4. LID = local item dependencies; ICC = item characteristic curve.

most common methods of handling LID is to introduce an item-level RE into the response kernel (e.g., Bradlow et al., 1999). Such an addition, however, invariably changes the interpretation of the parameters of an item, making its interpretation dependent on the specific RE of the cluster to which the item belongs. As a result, it also changes the interpretation of the weights of item attributes. A similar dilemma is also observed in the use of mixed models versus marginal models in the literature on clustered-data analysis (e.g., Fitzmaurice, Laird, & Ware, 2004, chap. 13).

To understand what is meant by the marginal interpretation of the item parameters, consider an example in which three binary items (Y_1, Y_2, Y_3) are known to form an item cluster. That is, there exists a nonzero correlation between each item pair within the cluster, given ability. For each item, let the item characteristic curve (ICC) be given by

$$P(Y_{ij} = 1|\theta_j) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)}, \tag{1}$$

where Y_{ij} denotes the item response of the jth person to the ith item, θ_j denotes the (latent) trait score of the person j, and b_i represents the item difficulty parameter of item i, $i = 1, 2, 3$. We assume that the values of b for the three items are, respectively, 0.2, 0, and –0.2. Therefore, for a person with $\theta = 0$ (person index omitted), the probability of a positive response to Item 1 is given by $P(Y_1 = 1|\theta = 0) = 0.4158$. Under the assumption of local independence, the probability of observing the pattern $Y_1 = 1, Y_2 = 1, Y_3 = 1$ for this person is the product of the three probabilities $P(Y_i = 1|\theta = 0)$, $i = 1, 2, 3$. However, under LID the joint probability of a specific pattern would be different from the value given by the product rule.

Table 1 shows how probabilities of response patterns change when LID, measured in terms of the pairwise conditional odds ratio, is present. In Table 1, the three marginal probabilities are constrained to the value given by equation (1) in the presence of LID. The term *marginal* is used here to distinguish between probability distributions such as $P(Y_1 = 1|\theta = 0)$, which we call marginal, and conditional distributions such as $P(Y_1 = 1|\theta = 0, Y_2 = 0, Y_3 = 0)$. The marginal probability of the item response for Item 1 is the sum of the probabilities of the four patterns (100), (110), (101), and (111), and it can be easily verified that under LID in Table 1, $P(Y_1 = 1|\theta = 0) = P^*(100) + P^*(110) + P^*(101) + P^*(111) = 0.4158$, which exactly equals the value of the item response function of the locally independent model given

Downloaded from http://apm.sagepub.com at Katholieke Universiteit Leuven on December 16, 2009

by equation (1): $P(Y_1 = 1|\theta = 0) = 0.4158$. An invariant marginal model requires that such agreement holds for all values of θ. It is also easy to verify that under local independence, $P(100) + P(110) + P(101) + P(111) = 0.4158$.

Thus, regardless of whether LID is present, and regardless of which subset of items is under consideration, the interpretation of the item parameters remains the same under the marginal model. The distinction between models that have invariant marginal interpretation and those that do not is a point of departure for the proposed LID-LLTM and other existing item-response models. The marginal property is not shared by LID models such as the RE testlet model. Nor is it shared by log-linear models.

In this article, the motivation for developing LID-LLTM is to properly model the item attributes within a measure of verbal aggression, after controlling for LID and other personal characteristics. The application uses data from a situation-response (S-R) inventory concerning verbal aggression. Several features can influence the probability of a person becoming verbally aggressive in a situation, some of which are situational, whereas others are related to personal characteristics (see, Infante & Rancer, 1996; Smits, De Boeck, & Vansteelandt, 2004). Each hypothetical situation can be regarded as a random sample drawn from a universe of real-life situations. Here, the marginal property of the item parameter becomes important because it allows the meaning of the person attribute parameters to be independent of the specific item cluster (situation) to which the item belongs. The dependency of different verbally aggressive behaviors within the same situation, after accounting for the situational features and individual dispositions toward aggressiveness, is considered as manifesting itself in the form of positive or negative associations between item pairs within a situation.

Method

Data

The participants were 316 first-year psychology students (73 men and 243 women; average age = 18.4, $SD = 1.2$). Three surveys (described below) were administered to these participants at the Catholic University (K.U.) of Leuven, Belgium, and participation in the study was a partial fulfillment of a requirement to participate in research. This data set is a subset of a comprehensive data set described in Smits et al. (2004).

The data set contains item responses collected from three instruments. The first instrument is a subset of an inventory concerning the presence/absence of two verbally aggressive behaviors (cursing and scolding) in four frustrating situations. Each of the four situations was presented to the participants and followed by two questions: (a) Would the participant curse (the word is translated from Dutch, and the Dutch word literally means that the action is not necessarily directed toward another party) when in that situation? and (b) Would the participant scold (often directed toward another party) when in that situation? Thus, in the data set there are, altogether, eight items (four situations times two questions per situation).

The four situation descriptions used in this article were taken from two existing situation-response inventories by Endler and Hunt (1968). They are situations with which students are familiar and (a) are likely to evoke anger, or (b) represent what are considered to

Downloaded from http://apm.sagepub.com at Katholieke Universiteit Leuven on December 16, 2009

Table 2
Four Situations in the Situation-Response Inventories and Percentage of Positive Response in Data Situation

	Item Covariate Behavior Type	Item Covariate Situation Type	Percentage of Positive Response
Flat tire	Curse	Material breakdown	70.9
	Scold	Material breakdown	47.2
Awakened by loud noises	Curse	Person	41.7
	Scold	Person	27.5
Grocery store clerk gives	Curse	Person	45.9
wrong information	Scold	Person	24.4
Broken typewriter	Curse	Material breakdown	70.6
	Scold	Material breakdown	45.6

be personally frustrating events. The four situations, described in Table 2, were translated into Dutch by two independent translators. Agreement across translators was used as a criterion to select the current translations.

The four situations can be subdivided into two types: (a) two situations in which a person is responsible for the frustrating event, and (b) two situations in which a material breakdown causes the frustration. By consequence, two situation-specific attributes can be identified: the type of situation, and a behavior-related attribute—whether the question is about cursing or about scolding.

To disguise our design and to counteract response tendencies, the two behaviors were mixed, for each situation, with 19 other behaviors, and the four situations were mixed with 11 other situations. In addition, the order of the items was randomized.

The collected data also contain person-specific features, which can influence the probability of becoming verbally aggressive, a behavior that is widely considered to be related to anger and its expression because anger is often conceptualized as the emotion that motivates aggression (e.g., Averill, 1983). The two measures, respectively, for anger and anger expression were (a) the Dutch adaptation of the State-Trait Anger Scale (Spielberger, Jacobs, Russell, & Crane, 1983), which is a measure of trait anger and (b) the Self-Expression and Control Scale (SECS; Van Elderen, Maes, Komproe, & van der Kamp, 1997), which contains the subscales Anger Out, Anger In, Anger In Control, and Anger Out of Control.

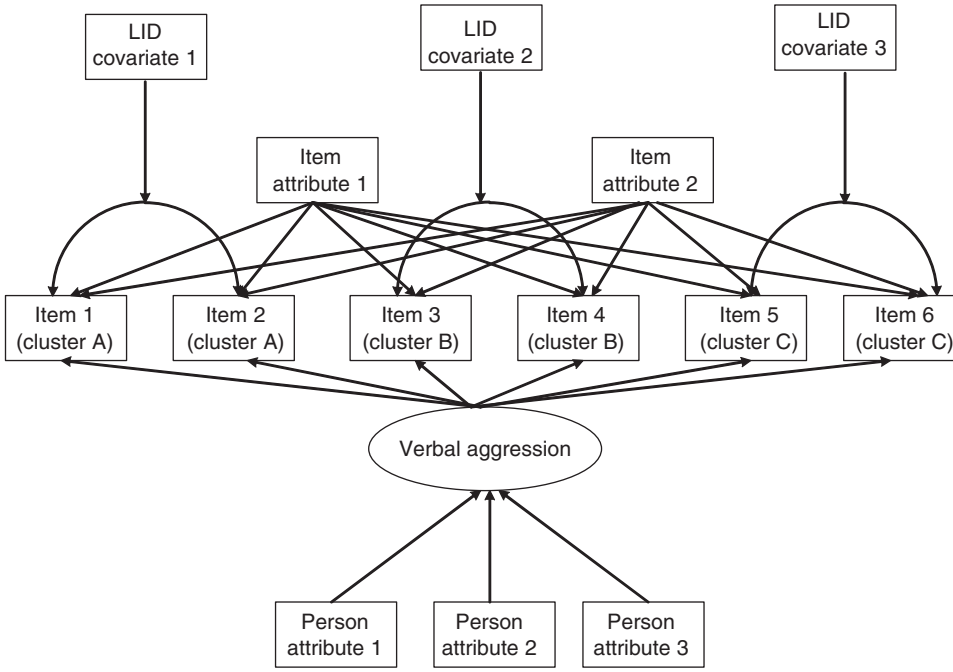
Model

A general formulation of LID-LLTM is presented followed by a description of the specific model used in the current application. The basic workhorse of LID-LLTM is the Rasch model, as exemplified by equation (1), but clearly other logistic models, such as the two-parameter logistic model (2PL), can also be accommodated without much modification of the method.

The overall structure of the model and its components—item attributes, person characteristics, and LID—are depicted in Figure 1, and each component is separately modeled as

Downloaded from http://apm.sagepub.com at Katholieke Universiteit Leuven on December 16, 2009

Figure 1
Schematic Display of the Structure of a Hypothetical Example of the LID-LLTM Model



Note: In this example, there are three item clusters (A, B, and C), each with two items, three person attributes, and one covariate for the LID component. The LID covariate may or may not be the same across clusters. LID-LLTM = locally item dependent linear logistic test model.

follows. First, the item attributes follow the linear logistic form of a fixed-effects model (Fischer, 1973):

$$b = X\beta, \tag{2}$$

where X is the design matrix specifying the structure of the items present in the test, and b, β are, respectively, the vector-of-difficulty parameters and the vector of nonrandom coefficients of item attributes. An intercept term β_0, which is common to all items, is included as part of β in equation (2). The marginal representation in equation (2) is critical for the LLTM because such representation allows the β coefficient, which is indicative of the extent to which each attribute affects item difficulty, to have a common and consistent interpretation across items and clusters.

Person covariates, which are contained in the matrix **Z**, are incorporated as a separate regression model for the latent trait. In other words, θ_j in equation (1) is conceptualized as an unobserved person-specific factor that can either be explained by other observable

Downloaded from http://apm.sagepub.com at Katholieke Universiteit Leuven on December 16, 2009

measurements or influenced by background characteristics such as demographics. Formally, such a relation can be expressed as

$$\theta_j = Z_j^T \gamma + e_j, \tag{3}$$

where Z_j is a vector of observed personal characteristics for person j, γ is a vector indicating the linear effects of each personal variable, and e_j is a random component that follows a normal distribution. To identify the scale of the measured trait, we assume that e_j follows a normal distribution with zero mean. Finally, the LID component can be modeled as a function of the latent trait and other covariates that are expected to contribute to LID. The magnitude of LID can be operationalized through a procedure called mixed parameterization (Barndorff-Nielsen, 1978), and eventually it can be quantified as the pairwise conditional odds ratios between item pairs within the same cluster (Ip, 2002; Ip et al., 2004). To avoid fitting unnecessarily complex models of association, all higher-order interactions in the LID model are set to zero.

The general formulation of the model for the conditional log-odds ratio between Item j and Item k, ω_{ik}, follows the form

$$\omega_{ik} = \lambda_0^{(ik)} + \lambda_1^{(ik)}\theta + \sum_{m=2}^{M} \lambda_m^{(ik)} w_m^{(ik)}, \tag{4}$$

where $\lambda_0^{(ik)}, \lambda_1^{(ik)}, \lambda_m^{(ik)}$ are item pair-specific coefficients, and $w_m^{(ik)}$ is an item pair-specific covariate. Here, the conditional log-odds ratio is a measure of residual association between item pairs, given θ, and all other item responses within the same cluster, typically all set at the value of 0. The general form expressed in equation (4) allows the statistical testing of hypotheses such as whether LID is related to the latent trait or other covariates. Accordingly, it can be tested whether the dependency between the item pairs of cursing and scolding in a specific situation can be explained by the person's overall tendency toward being verbally aggressive.

The setup in equation (4) also allows the testing of the hypothesis that anger measures such as Anger Out of Control have an effect on the divergence of cursing (not toward anyone) and scolding (toward others). When sample sizes are small, one may want to assume some form of constancy of LID over the entire cluster and model LID as a function of cluster characteristics rather than of characteristics of individual item pairs. For example, a simple model is to assume constancy in conditional log-odds ratios across all item pairs:

$$\omega_{ik} = \lambda_0. \tag{5}$$

In other words, λ_0 represents the uniform LID between any pair of items within a testlet, at any level of θ. An even stronger assumption would be to have constancy of LID over all item clusters (situations in the current example).

The focus is now on several LID-LLTM models by comparing their goodness of fit to the data. Because there exists a rather broad range of models that all have good face validity, the model that was representative and consistent with psychological theory was selected.

Downloaded from http://apm.sagepub.com at Katholieke Universiteit Leuven on December 16, 2009

The model was used as the theory-based model and is referred to as the comparison model, against which all other models were compared. The comparison model used the following input variables: for item attributes—material breakdown versus person, curse versus scold; for person attribute—Anger Out; and for LID—a different odds ratio for each situation, but without covariates. The comparison model was then compared with other nested models.

The following competing families of models were fitted and compared with the comparison model:

1. Model A (item component): Added and deleted item-attribute variables, especially adding an interaction term for Material Breakdown and Curse, as we expect respondents to curse more when Material Breakdown is involved.
2. Model B (LID component): Added verbal aggressiveness and other predictors such as Anger Out of Control and gender to the LID regression component.
3. Model C (person component): Used different anger measures (e.g., Trait Anger) as person covariates.
4. Model D: Used a locally independent model.

The G^2 statistic (–2 × log likelihood) was used to assess how well the model fits the data and to guide the selection of the final model. A lower G^2 value implies a better fit. Under general regularity conditions, the change in G^2 follows as a chi-squared distribution with $n - 1 - p$ degrees of freedom, where p is the number of additional parameters. The statistical significance of the change in G^2 is often used for determining whether a more complex model (one with more parameters) is better than a reference model. However, as many authors have pointed out (e.g., Wang, Chen, & Wilson, 2005), experience has shown that in many cases the G^2 statistic gravitates toward more complex models. Although such models generally give better fit to the data in terms of G^2, it is often desirable to select models that are parsimonious and meaningful but that still provide a reasonable overall fit. Following Wang et al. (2005), we examined the goodness of fit through both the change in G^2 and the change in G^2 per parameter fitted in competing nested models.

Estimation

An expectation–maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) was used to estimate the model. The EM algorithm is a statistical method for making statistical inferences in the presence of missing data. Following the work of Ip (2002) for the EM estimation of marginal LID models, we developed additional features for estimating the proposed model, which include the estimation of regression coefficients for both item and person attributes. Because the basic algorithm for estimating LID models in Ip (2002) is based on iterative estimation of one set of parameters given another, the modularity of the algorithm makes the incorporation of the regression models into the item and the person components relatively straightforward.

In brief, the proposed EM algorithm extends the EM solution for IRT estimation proposed by Bock and Aitkin (1981). The algorithm consists of the expectation (E) and maximization (M) steps. In the E step, the latent variable θ is treated as missing data, and $P(\theta|y, \beta, \lambda, \gamma)$, the posterior density given some provisional values of the parameters, is

Downloaded from http://apm.sagepub.com at Katholieke Universiteit Leuven on December 16, 2009

computed. The M step, on the other hand, solves the log likelihood equation for the parameters (β, λ, γ), and the procedure involves the posterior density obtained in the E step. To maintain the invariance of the marginal ICC of item response in LLTM, the E step relies on an iterative proportional fitting (IPF) algorithm (Deming & Stephan, 1940; for computation in an IRT context, see Ip, 2002) to fit joint distributions of clustered item responses at quadrature points of the latent trait in such a way that the marginal distribution of each item follows a specific ICC for a given LID structure. The computer program for estimation was developed in Splus and can be obtained from the first author.

Parameter Recovery

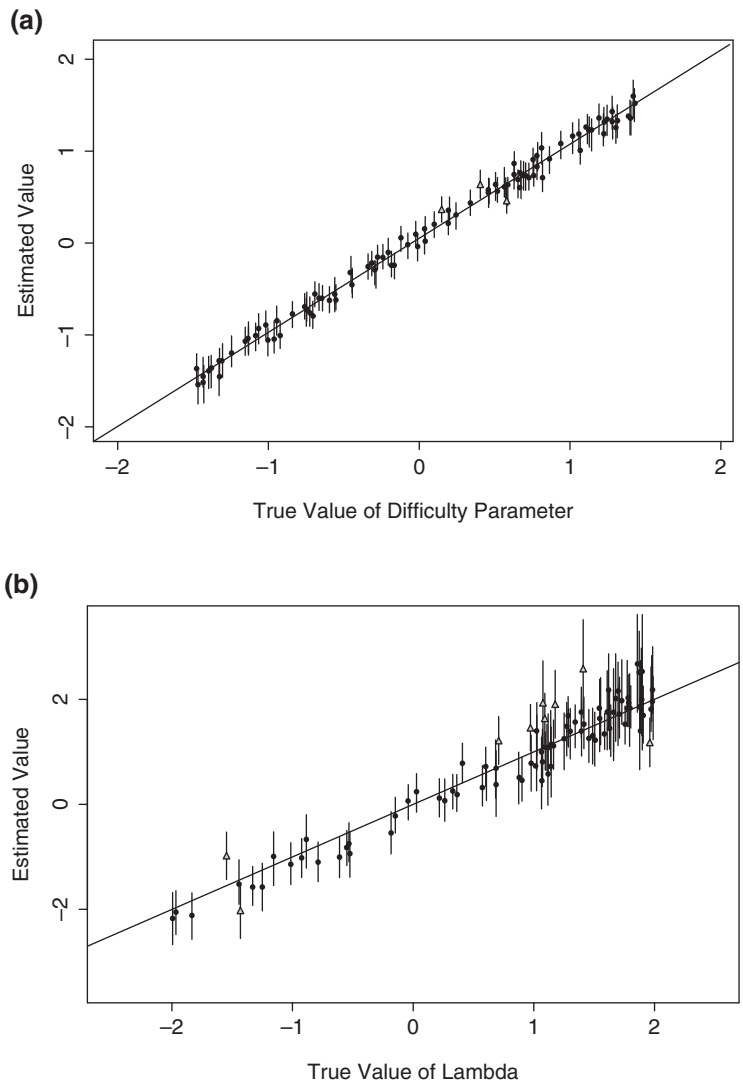
To assess the capacity of the above estimating algorithm for recovering correct parameter values, two small simulation experiments were conducted. The first simulation experiment aimed at focusing on the recovery of parameters in the LID model, whereas the second experiment aimed to provide evidence of parameter recovery for the LLTM component, using simulated data that had a structure similar to that of the Anger survey data. The metric for evaluating parameter recovery is the coverage probability of the 95% confidence limits—that is, the percentage of times that the estimated confidence limits produced by the proposed procedure include the true parameter.

In the first experiment, a total of 20 items were divided into 10 clusters in such a way that the conditional odds ratio followed the model given by equation (4). It was assumed that for cluster $k, k = 1, \cdots, 10$, the odds ratio between the item pair within each cluster followed the model $\omega_k^{(12)} = \lambda_{0k}^{(12)} + \lambda_{1k}^{(12)}\theta$, with λ_{0k} generated from a uniform distribution with end points (–2, 2), and λ_{1k} generated from a uniform distribution with end points (1, 2). The difficulty parameter was generated from a uniform distribution with end points (–1.5, 1.5), and the individual latent-trait parameter θ_j was assumed to follow a standard normal distribution. One thousand individual θ_j's were generated for this simulation experiment. The data set was then analyzed using the LID model. The above procedure was replicated five times to produce results for five sets of simulated responses. Figure 2a and 2b show the coverage of the confidence limits for the true item parameters and the LID parameters, respectively. The coverage probability for item parameters was 97%, whereas that for the LID parameters was 91%. The results show that the proposed procedure recovers both the item and LID parameters quite well. Compared with the standard error of the item parameters, the standard errors for the LID parameters were much higher, suggesting that the LID parameters, the weight parameters for latent trait in particular, were not as accurately measured. For the item parameters, the mean deviation (mean difference between true and estimated values, a measure of potential bias) and the root mean square error (RMSE) were, respectively, 0.004 and 0.089, whereas for the LID parameters they were, respectively, –0.023 and 0.24.

In the second experiment, 50 data sets, each structured similarly to the verbal aggression data (four situations, with two clustered items within each situation) were simulated in accordance with the following conditions. For each simulated data set, two item-attribute β parameters were generated from a uniform distribution with end points (–1.5, 1.5) and with the intercept term β_0 set to 0; the item-level design matrix values of X were randomly filled

Downloaded from http://apm.sagepub.com at Katholieke Universiteit Leuven on December 16, 2009

Figure 2
Graph of Point Estimate and 95% Confidence Interval of (a) Difficulty Parameter and (b) LID Parameters λ_0 and λ_1 in Experiment 1

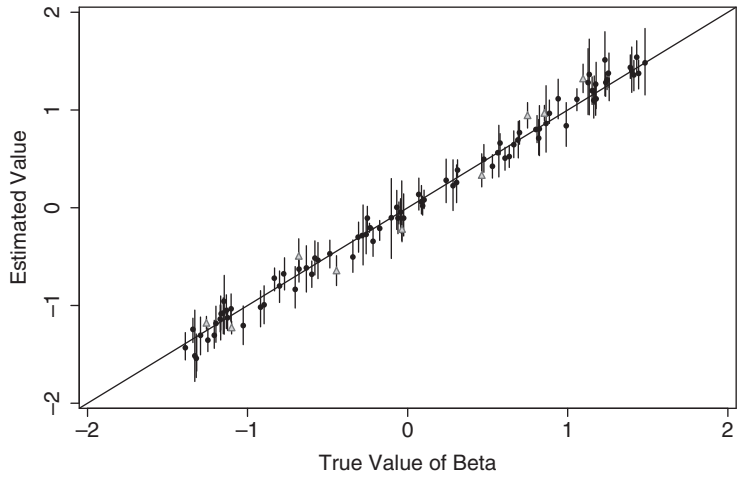


Note: Points covered and not covered by the 95% confidence intervals are represented by dots and triangles, respectively. LID = local item dependencies.

with 0 and 1; the person covariate was a randomly sampled integer in the same range as in the Anger Out scale, which is (11, 39); the person parameter γ was generated from a uniform distribution with end points (0, 0.005); and the situation-specific constant LID parameter $\omega = \lambda_0$ was generated from a uniform distribution with end points (–2, 2) for each situation. For each data set, we generated $n = 500$ participants with θ sampled from a standard normal distribution.

Downloaded from http://apm.sagepub.com at Katholieke Universiteit Leuven on December 16, 2009

Figure 3
Graph of Point Estimate and 95% Confidence Interval of the LLTM Parameters β in Simulated Experiment 2



Note: Points covered and not covered by the 95% confidence intervals are represented by dots and triangles, respectively. LLTM = linear logistic test model.

This simulation experiment suggested that the 95% confidence intervals for the LLTM parameter β, the person parameter γ, and the LID parameter λ_0 all cover the population values rather well. For β, of 2 (attributes) × 50 (data sets) = 100 parameter estimates, nine were not covered, with five in one direction and four in the other. Figure 3 shows parameter recovery for β. The mean deviation between β and the estimate $\hat{\beta}$ was –0.001, and the RMSE for β was 0.075. For γ and λ_0, the proportions of estimates not covered by the 95% confidence interval were, respectively, 4% and 9%. For the sake of parsimony, only the graph for parameter recovery for β in Figure 3 is included.

Results From Real Data Analysis

The results from the evaluation of competing models are reported followed by the results from the best-fitted model. For models of person covariates (Model C), because the correlations between the various anger measures were rather high, in each competing model only a single anger measure was included as a predictor to avoid the problem of multicollinearity. The range of correlations between the five anger measures—Anger Trait, Anger Out, Anger Out of Control, Anger In, Anger in Control—have a range between 0.3 and 0.8, if the correlation between Anger Trait and Anger In (correlation = −.05) is excluded. Hypothesis tests showed that except for the measure Anger Out none of the anger measures was statistically significant at the 0.10 level.

Downloaded from http://apm.sagepub.com at Katholieke Universiteit Leuven on December 16, 2009

Table 3
Parameter Point Estimates (Standard Errors) From the LID-LLTM Analysis of Verbally Aggressive Behaviors

	Comparison	Model A	Model B[a]	Model C	Model D
β[b] (Item)					
Material[c]	1.09	0.59	1.09	1.09	1.12
	(0.38)*	(0.48)	(0.38)*	(0.38)*	(0.40)*
Curse[c]	1.65	1.15	1.64	1.65	1.69
	(0.38)*	(0.47)*	(0.38)*	(0.38)*	(0.40)*
Material × Curse		1.01			
		(0.68)			
γ (person)					
Anger out	0.0034	0.0025	0.0041		0.0047
	(0.0018)	(0.0019)	(0.0019)*		(0.0020)*
Anger trait				0.0041	
				(0.0032)	
Variance σ^2_θ	1.85	1.83	1.83	1.85	2.10
	(0.25)	(0.25)	(0.25)	(0.25)	(0.26)
λ (LID)					
Situation 1	0.15	0.15	–0.72	0.05	0.15
	(0.42)	(0.42)	(0.51)	(0.43)	(0.42)
Situation 2	1.98	1.99	0.22	1.89	1.98
	(0.40)*	(0.40)*	(0.40)	(0.40)*	(0.40)*
Situation 3	1.58	1.58	–0.04	1.57	1.57
	(0.37)*	(0.38)*	(0.37)	(0.37)*	(0.37)*
Situation 4	1.72	1.72	–0.77	1.61	1.72
	(0.38)*	(0.38)*	(0.47)	(0.40)*	(0.38)*
Number of parameters	9	10	12	9	5
G^2	3,053.6	3,053.0	3,038.8	3,056.2	3,245.2
$\Delta G^2/\Delta df$[d]	–0.6	–3.7			–47.9

a. The first column in LID refers to the coefficient of latent trait λ_{1k}, the second to the intercept λ_{0k}.
b. Intercept β_0 not shown.
c. Person and Scold are, respectively, the reference categories.
d. change in G^2 per degree of freedom as compared to the comparison model.
LID-LLTM = locally item dependent linear logistic test model; LID = local item dependencies.
*$p < .05$.

Space limitations preclude the comprehensive reporting of all of the tested families of models in (A) to (D). Table 3 reports the results from the comparison model and the most competitive model from each category: for (Model A), addition of Material Breakdown by Curse interaction; for (Model B), verbal aggressiveness (latent trait) as covariate; and for (Model C), Trait Anger as covariate. The locally independent Model D is also included for comparison.

From Table 3, it can be seen that the comparison model used nine parameters and had a G^2 value of 3,053.6. Among the competing models, Model B had the best G^2 value but used the greatest number of parameters. There were 12 parameters, but the G^2 change per parameter (3.7) over the comparison model was not substantial. However, the G^2 difference per

Downloaded from http://apm.sagepub.com at Katholieke Universiteit Leuven on December 16, 2009

parameter between the comparison model and the locally independent model was sizeable (47.9). Also, adding an interaction term β for Material Breakdown $\times$ Curse did not significantly improve the fit. Using Anger Out in the comparison model leads to the best fit among models using different anger and anger-expression measures. In summary, based on consideration of the preferred G^2 per parameter criterion, as suggested by Wang et al. (2005), the comparison model is the model of choice. Subsequent interpretation of the parameters will be based on this model.

Except for the association parameter λ for situation 1 (flat tire) and Anger Out, the parameter estimates in the comparison models are all significant at the 0.05 level. For the item-attribute parameter β, the reference categories are, respectively, Person and Scold. Thus, both Material Breakdown and Curse items will elicit significantly more aggressive verbal behavior. For example, at the value of $\theta = 0$, if an item concerns material breakdown it has a probability of 0.73 of eliciting verbally aggressive behavior, compared with a probability of 0.5 when an item concerns a person. This result is consistent with the prevalence of the behaviors (observed frequencies in the last column of Table 2). The anger-expression measure Anger Out is modestly related to the latent trait of verbal aggressiveness ($p = .06$). For the LID component, all association parameters λ_0 are positive, showing a positive relationship between the pair of verbally aggressive behaviors. The results from Model B suggest that there is little or no relationship between this association and latent trait (first column, Model B in Table 3). Additional analyses (not reported) also indicated that other anger measures and gender do not have a significant relationship with the extent of association between the two behaviors.

The computing time for the various LID models ranged from 3.6 to 6.5 minutes on a shared, networked SunFire V880 server with a central processing unit (CPU) speed of 8,900 MHz and 16 GB of RAM. The largest computational overheads were derived from fitting the ICCs using IPF.

Discussion

The findings from the previous section suggest that there is strong evidence to support the claim that LID is present within a situation. Except for the first situation, the odds ratios lie in the range from $\exp(1.58) = 4.86$ to $\exp(1.98) = 7.24$, which suggests that inward and outward verbal aggressive behaviors are highly correlated, even after controlling for individual differences. From a modeling perspective, this level of LID indicates that ignoring LID may lead to biased estimates of item parameters (Tuerlinckx & De Boeck, 2001) and to biased ability estimates and their associated standard errors (Bradlow et al., 1999; Ip, 2000). The findings from the present study also suggest that LID between the behaviors does not depend on the level of verbal aggressiveness.

The item attribute effects, breakdown and cursing, are both strong. However, no interaction effect between the two attributes was observed—whether the situation involves materials breakdown and the type of verbal aggressive behavior. This was somewhat surprising because it was expected that respondents would curse more when there was material breakdown since there is no one to blame. Perhaps this can be explained by the fact

Downloaded from http://apm.sagepub.com at Katholieke Universiteit Leuven on December 16, 2009

that cursing, the nonoutward-directed form of verbal aggressive behavior, occurs more frequently and regardless of situation. From results regarding the person submodel, it can be seen that the person covariate Anger Out ($p = .06$) does not predict verbal aggressiveness very well. This is an indication that the construct, verbal aggressiveness, has only moderate overlap with other anger and anger-expression measures.

References

Averill, J. R. (1983). Studies on anger and aggression: Implications for theories of emotion. *American Psychologist*, *38*, 1145-1160.

Barndorff-Nielsen, O. E. (1978). *Information and exponential families in statistical theory.* New York: John Wiley.

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

Bradlow, E., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153-168.

Butter, R., De Boeck, P., & Verhelst, N. D. (1998). An item response model with internal restrictions on item difficulty. *Psychometrika*, *63*, 1-17.

Deming, W. E., & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, *11*, 427-444.

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

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

Endler, N. S., & Hunt, J. M. (1968). S-R inventories of hostility and comparisons of the proportions of variance from persons, behaviors, and situations for hostility and anxiousness. *Journal of Personality and Social Psychology*, *9*, 309-315.

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

Fischer, G. H. (1995). The linear logistic test model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundation, recent developments, and applications* (pp. 131-156). New York: Springer-Verlag.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis.* Hoboken, NJ: Wiley.

Gibbons, R. D., & Hedekker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*, 423-436.

Glas, C. A. W. (2005). A review of *Explanatory item response models: A generalized linear and nonlinear approach* (P. de Boeck & M. Wilson, Eds.). *Journal of Educational Measurement, 42*, 303-307.

Hoskens, M., & De Boeck, P. (1997). A parametric model for local dependence among test items. *Psychological Methods*, *2*, 261-277.

Infante, D. A., & Rancer, A. S. (1996). Argumentativeness and verbal aggressiveness: A review of recent theory and research. In B. R. Burleson & A. W. Kunkel (Eds.), *Communication yearbook* (pp. 319-352). Thousand Oaks, CA: Sage.

Ip, E. H. (2000). Adjusting for information inflation due to local dependency in moderately large item clusters. *Psychometrika*, *65*, 73-91.

Ip, E. H. (2002). Locally dependent latent trait model and the Dutch identity revisited. *Psychometrika*, *67*, 367-386.

Ip, E. H., Wang, Y., De Boeck, P., & Meulders, M. (2004). Locally dependent latent trait model for polytomous responses with application to inventory of hostility. *Psychometrika, 69*, 191-216.

Jannarone, R. J. (1986). Conjunctive item response theory kernels. *Psychometrika*, *51*, 357-373.

Janssen, R., & De Boeck, P. (1997). Psychometric modeling of componentially designed synonym tasks. *Applied Psychological Measurement*, *21*, 37-50.

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

Downloaded from http://apm.sagepub.com at Katholieke Universiteit Leuven on December 16, 2009

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Mahwah, NJ: Erlbaum.

Mislevy, R. J., & Sheehan, K. M. (1989). The role of collateral information about examinees in item parameter estimation. *Psychometrika*, *54*, 661-679.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danish Institute for Educational Research.

Rijmen, F., & De Boeck, P. (2002). The random weights linear logistic test model. *Applied Psychological Measurement, 26*, 271-285.

Rosenbaum, P. R. (1988). Item bundles. *Psychometrika*, *53*, 349-359.

Scott, S., & Ip, E. H. (2002). Empirical Bayes and item clustering effects in latent variable hierarchical models: A case study from the National Assessment of Educational Progress. *Journal of the American Statistical Association*, *97*, 409-419.

Smits, D. J. M., De Boeck, P., & Vansteelandt, K. (2004). The inhibition of verbally aggressive behaviour. *European Journal of Personality*, *18*, 537-555.

Spielberger, C. D., Jacobs, G. H., Russell, S. F., & Crane, R. S. (1983). Assessment of anger: The state-trait anger scale. In J. N. Butcher & C. D. Spielberger (Eds.), *Advances in personality assessment* (Vol. 2, pp. 20-36). Hillsdale, NJ: Erlbaum.

Tuerlinckx, F., & De Boeck, P. (2001). The effect of ignoring item interactions on the estimated discrimination parameters. *Psychological Methods, 6*, 181-195.

Van Elderen, T., Maes, S., Komproe, I., & van der Kamp, L. (1997). The development of an anger expression and control scale. *British Journal of Health Psychology*, *2*, 269-281.

Wainer, H., Bradlow, E., & Wang, X. (2007). *Testlet response theory and its applications.* New York: Cambridge University Press.

Wang, W., Cheng, Y., & Wilson, M. (2005). Local item dependence for items across tests connected by common stimuli. *Educational and Psychological Measurement*, *65*, 5-27.

Wang, W., & Wilson, M. (2005). Exploring local item dependence using a facet random-effects facet model. *Applied Psychological Measurement*, *29*, 296-318.

Wilson, M. (1989). Empirical examination of a learning hierarchy using an item response theory model. *Journal of Experimental Education*, *57*, 357-371.

Zinderman, A. H. (1997). Response models with manifest predictors. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 245-256). New York: Springer.

Acknowledgment

This research was supported by National Science Foundation grant SES-0417349 awarded to Edward H. Ip.

Downloaded from http://apm.sagepub.com at Katholieke Universiteit Leuven on December 16, 2009