Psychological Methods
2001, Vol. 6, No. 2, 181–195

Copyright 2001 by the American Psychological Association, Inc.
1082-989X/01/$5.00 DOI: 10.1037//1082-989X.6.2.181

The Effect of Ignoring Item Interactions on the Estimated Discrimination Parameters in Item Response Theory

Francis Tuerlinckx and Paul De Boeck
University of Leuven

Most item response theory models assume conditional independence, and it is known that interactions between items affect the estimated item discrimination. In this article, this effect is further investigated from a theoretical perspective and by means of simulation studies. To this end, a parametric model for item interactions is introduced. Next, it is shown that ignoring a positive interaction results in an overestimation of the discrimination parameter in the two-parameter logistic model (2PLM), whereas ignoring a negative interaction leads to an underestimation of the parameter. Furthermore, it is demonstrated that in some cases the item characteristic curves of the 2PLM and of an item involved in an interaction are quite similar, indicating that the 2PLM can provide a good fit to data with interactions.

When measurement instruments for psychological assessment are constructed, it is often assumed that the only reason items are related to each other is that they tap the same underlying unobservable construct (or constructs). However, previous research has shown that contextual factors such as item wording, question order, and so on (Knowles & Condon, 2000; Schwarz, 1999) may exert an additional influence on the item response. In this article, we focus on the situation in which the responses to one or more items have an influence on the response to some other item, that is, on cases in which items interact with one another.

Examples of situations in which item interactions arise are legion. For instance, the solution of one ability item might indicate the correct answer to another such item (Hoskens & De Boeck, 1997). Yet another example occurs when two items in a reading test refer to the same part of the text (Yen, 1993). As a final example, Knowles and Condon (2000) have reported that grouping similar items in a test might lead to a higher intercorrelation among their responses.

The occurrence of item interactions causes difficulties for the statistical analysis of the measurements. We show that it might be difficult to detect those item interactions and that item interactions can lead to incorrect conclusions about the quality of the items and the test. These effects are shown to arise with models from item response theory.

Item response theory (IRT) refers to a collection of statistical models that link characteristics of the item and the person, on the one hand, to the probability that the person will emit a particular response to the item, on the other hand. In contrast to classical test theory, the unit of analysis in IRT is not the total test score but the individual item; therefore, it is more suitable for a detailed analysis of a set of questions (for an introduction to IRT, see Embretson & Reise, 2000 and Hambleton, Swaminathan, & Rogers, 1991).

Items are related to each other if they measure the same construct, and IRT models make use of such interdependencies by inferring latent traits to account for the covariation between items. We restrict our attention to IRT models that assume a single continuous latent trait. The central model considered is the two-parameter logistic model (2PLM; Birnbaum, 1968) which has the following formula:

Francis Tuerlinckx and Paul De Boeck, Department of Psychology, University of Leuven, Leuven, Belgium.

This research was supported by Grant BOF-GOA/02/2000 from the University of Leuven. Francis Tuerlinckx was a Research Assistant of the Fund for Scientific Research—Flanders (Belgium). We thank Machteld Hoskens, Peter Kuppens, Michel Meulders, Jeremy Pacht, Frank Rijmen, and Norman Verhelst for their critical reading and useful suggestions.

Correspondence concerning this article should be addressed to Francis Tuerlinckx, Department of Psychology, University of Leuven, Tiensestraat 102, B-3000 Leuven, Belgium. Electronic mail may be sent to francis.tuerlinckx@psy.kuleuven.ac.be.

$$\Pr(X_i = 1|\theta) = \frac{\exp(\alpha_i(\theta - \beta_i))}{1 + \exp(\alpha_i(\theta - \beta_i))}. \quad (1)$$

In Equation 1, X_i is the random variable denoting the response of a person on item i, θ is the position of the person on the latent trait, β_i is the position of item i on the same continuum, and α_i is the item discrimination. The item discrimination quantifies the strength of the relation between the latent trait and the item. Note that if the item discriminations do not differ over items, the model reduces to the common one-parameter Rasch model.

It will be assumed that we are dealing with ability items. This means that θ may be considered as the latent ability of the examinee; β_i is the difficulty parameter of the item, and X_i is a random variable that takes a value of 1 if item i is solved correctly and a value of 0 otherwise. However, the line of argument presented in this article is also valid for other types of items, such as personality or attitude items.

The application of the 2PLM relies on three important assumptions (see, e.g., Holland & Rosenbaum, 1986; Rosenbaum, 1984). The first of these is *monotonicity*, which means that the probability of giving a correct response increases with the ability of the respondent. Second, it is assumed that the set of questions is unidimensional such that only one latent trait is needed to account for the covariation between the item responses. The third important assumption is called *conditional independence* (CI), and it implies that if the latent trait score of the person is known to be θ, then all association between different items should disappear. This is formalized as follows:

$$\Pr(X_1 = x_1, \ldots, X_K = x_K|\theta) = \prod_{i=1}^{K} \Pr(X_i = x_i|\theta), \quad (2)$$

where the lowercase x_is are 0 (for incorrect responses) or 1 (for correct responses).

When the equality in Equation 2 does not hold, the assumption of CI is violated. In that case, there are local item dependencies (LIDs) because, after conditioning on the latent trait, some nonnegligible (positive or negative) covariance remains between the items. However, LID is a broad and general term, and it may have several causes. We distinguish between two important sources of LID for IRT models. The first refers to more than one underlying dimension as a source of LID, and the second refers to a manifest variable as a source of LID.

The first cause of LIDs is *multidimensionality*, which means that at least one other latent variable plays a part in the response process. Thus, a single latent trait cannot account for all covariation between the item responses: After conditioning on that single latent trait, some covariation remains between the items.

The second cause of LIDs is that responses on some item (or items) may have consequences for the response probabilities on other items. This type of violation is called *item interaction*, and the variable that affects the responses is itself an item response and hence is a manifest variable.

In this article, we focus specifically on item interaction as a cause of LIDs.[1] To study item interactions, we partition the total item set into at least two subsets. The first subset contains the items that do not interact and that meet the traditional IRT assumption of CI, so that the 2PLM (Equation 1) can be applied to this subset. The second subset contains a set of interacting items, but the subset as a whole is conditionally independent of the first subset (and any other subsets). Hence, there is no CI within this subset, but there is CI between subsets. If there are other subsets of interacting items, they are grouped in additional subsets. Again, there is no CI within each of these new subsets, but there is CI between subsets. In this article, the subsets that show LIDs are modeled as if they were polytomous items. In addition, in accordance with Wainer and Kiely (1987), we refer to such a subset of items, which show LIDs and are modeled together, as a *testlet*.

For the items that interact, an explicit unidimensional model for item interaction must be defined, incorporating parameters to account for the interaction strength. Throughout the article, we show in various ways that when the 2PLM is used to analyze data that contain interactions, the estimated discrimination parameter is biased. Moreover, we show that the

[1] Although we are concerned with item interactions, this does not mean that multidimensionality as a cause of LIDs is conjectured as unimportant. An approach similar to the one taken in this article for investigating the distorting effect of item interactions on the estimated item discrimination could be taken starting from a model that incorporates multidimensionality. It then also makes sense to derive what the consequences are of neglecting multidimensionality. However, the to-be-presented interaction model is not properly parameterized to handle multidimensionality, so that our results cannot be directly generalized to that case.

2PLM is liable to be accepted as a well-fitting model, even when the true model is one that contains item interactions.

The research is focused on item discrimination because it is considered an important parameter in judging the quality of an item. If items are more strongly discriminating, they are believed to be better indicators of the latent trait. Moreover, the item discrimination has an influence on other features used to assess the quality of (a set of) items such as item and test information and the standard error of the estimated θ. The effect of ignored interactions on the estimated β_i is examined only very briefly, and we do not consider the influence on the estimated θ. We return to these issues in the Discussion section.

The structure of the remainder of the article is as follows. First, we provide a short overview of the literature on the topic. Second, we propose a parametric model for item interaction and briefly discuss its characteristics. Third, we show mathematically that ignoring the interaction affects the estimated discrimination parameter. Fourth, we investigate the concordance between the *item characteristic curve* (ICC, i.e., the curve of the probability of a correct response as a function of θ) of an item that is involved in an interaction and the ICC of a 2PLM item. Fifth, a series of simulation studies is performed to explore the influence of several factors, such as the number of items and number of interactions in the same data set, on the effect of the ignored interaction. Finally, we present our conclusions and discuss their implications.

For two reasons, the present article is restricted to the case of interactions between two items. First, these are the simplest and most probable forms of interaction (see, e.g., Hoskens & De Boeck, 1997). Second, interactions between more than two items are much more complex and therefore more difficult to interpret. Although our conclusions can be extended to interactions between more than two items, this is not shown in this article.

We do not want to suggest that differences in discrimination across items are always due to an ignored interaction. Instead, we suggest that it is useful to check for item interactions when the discriminations differ so that item interaction can be excluded as a possible source of these differences.

Previous Studies

Several studies can be found in the literature in which the effect of LIDs on the estimated parameters of the 2PLM is reported. In this section, a brief overview of the existing literature is given. In most of the studies, LIDs as such are investigated, and, because LIDs incorporate ignored item interactions, the results are also relevant to this study.

Thissen, Steinberg, and Mooney (1989) have noted that when there are positive LIDs that are not incorporated in a model, the test information for most θ values is substantially overestimated. They evaluated this overestimation by comparing the test information of a model fitted for the separate items against a model that is fitted for recoded items, taking into account expected LIDs. For the recoding, items that could be expected to show LIDs were grouped. For each group of items, one polytomously scored "superitem" was constructed, with response categories representing the number of correct responses to all items in the group; this is called the *score-based approach.* This superitem was defined on a testlet, as referred to above. Analyzing test data in this way can account for LIDs in a statistical sense. In a follow-up article, Sireci, Thissen, and Wainer (1991) have noted that, as a consequence of neglected LIDs, the *marginal reliability* (defined as the reliability of θ integrated over the distribution of θ) of the estimated person parameters is overestimated when the (wrong) model with the assumption of CI is used on a data set with positive LIDs.

Yen (1993) has shown that the test information for the latent trait estimates may be seriously overestimated if LIDs are neglected. Moreover, Yen claims that when discrimination parameters are estimated from data that show LIDs, these parameters tend to be higher for those items that have positive LIDs. Furthermore, she predicts that the standard errors of measurement of the estimated latent trait values are underestimated (of course, this is related to an overestimation of the test information). Like Sireci et al. (1991), Yen compared the fit of a model for separate items to that of a model for polytomous items, the latter constructed by combining those individual items that show LIDs. Steinberg and Thissen (1996) have also noted that positive LIDs may create large estimated discrimination parameters.

Chen and Thissen (1997) have introduced the concept of surface local dependence to refer to two items that are likely to have the same responses because both of them are not answered (e.g., because the test is long) or because they are very similar in content. Chen and Thissen have noted that if the surface local dependence increases and is not modeled, the esti-

mated discrimination parameter of the two items also increases.[2]

A Parametric Model for Item Interactions

In general, there are two ways to model item interactions in an IRT model (Hoskens & De Boeck, 1997; Wilson & Adams, 1995). One method is the aforementioned score-based approach. In the second approach, which is item-based, the categories of a polytomous item are each treated as a possible response pattern (and not just the sum scores) for the group of interacting items. The score-based and item-based approaches have in common that they analyze the interacting items (the testlet) with some kind of polytomous model. However, in the item-based approach, the parameters refer to the individual items and to specific interactions between items. In this way, no information about the items and interactions is lost. A drawback of the score-based approach is that the sum score is analyzed, so that the parameters are not interpretable in terms of phenomena that relate to specific items. In the same vein, using only the sum score results in a loss of information.

Several authors (Hoskens & De Boeck, 1995, 1997; Jannerone, 1986; Kelderman, 1984; Wilson & Adams, 1995) have described item-based parametric psychometric models for item interaction. In this article, we use the interaction model of Kelderman (1984) and Hoskens and De Boeck (1997). For simplicity, the model is applied to the case of two binary scored items, denoted by 1 and 2. The two responses to the items can be seen as the realization of a bivariate random variable (X_1, X_2), and the probability for a particular realization (x_1, x_2) is:

$$\Pr(X_1 = x_1, X_2 = x_2|\theta) = \frac{\exp(x_1\alpha_1(\theta - \beta_1) + x_2\alpha_2(\theta - \beta_2) + x_1x_2(-\beta_{12}))}{1 + \sum_{j=1}^{2}\exp(\alpha_j(\theta - \beta)) + \exp(\alpha_1(\theta - \beta_1) + \alpha_2(\theta - \beta_2) - \beta_{12})}, \quad (3)$$

in which β_{12} is the parameter that quantifies the interaction between items 1 and 2. Hoskens and De Boeck (1997) call this model the *constant interaction model*. The interaction is constant (on the logit scale) because it is equal in direction and magnitude for all subjects, independent of the latent trait value. Hoskens and De Boeck (1995, 1997) also describe a dimension-dependent interaction model (with the interaction being dependent on the latent trait value of the person), and they apply both types of model to the data, finding meaningful interpretations of the interactions.

In the constant interaction model, $\beta_{12} < 0$ means that there is a positive interaction between the two items and, hence, that the probability of a correct response to both items is larger than in the case in which $\beta_{12} = 0$ (all other things being equal). The reverse happens for $\beta_{12} > 0$, because in that case solving both items is more difficult than the difficulty parameters of the separate items would predict. If $\beta_{12} = 0$, Equation 3 simplifies to a product of two 2PLMs with parameters (α_1, β_1) and (α_2, β_2), which corresponds to the case of CI.

In the simulation study that is described below, values are assigned to the item parameters. To obtain an idea of the size of these values, it is important to know something about the population distribution of θ. Usually, a standard normal distribution for θ is assumed such that approximately 95% of the individuals fall between −2 and 2. Because the individual difficulty parameters β_i are located on the same scale as θ, values that range from −2 to 2 are realistic.

The metric of the interaction parameter β_{12} is the same as the one for the individual difficulty parameters β_i. This can be seen by computing the conditional probability $\Pr(X_2 = 1|X_1 = 1)$ from Equation 3, which has the same structure as the formula in Equation 1 except that the difficulty parameter is now equal to $(\beta_2 + \beta_{12})$. Hence, values between −2 and 2 also seem realistic for the interaction parameter. Values in this range have been found in empirical studies (Hoskens & De Boeck, 1997).

Next, we compute two quantities to obtain an idea of the effect of the interaction parameter. First, the expected value of a correlation between two item re-

[2] Bradlow, Wainer, and Wang (1999) have presented a general random-effects model for modeling dependencies in a testlet, in which the probability of a correct response is influenced by a person-specific testlet factor. However, in contrast to the empirical results we discussed, their model predicts that a larger dependency leads to an underestimation of the discrimination parameter when the dependency is not taken into account. When a CI model is applied to data from their model, the random (person-specific) testlet effect is considered as noise. This additional source of noise leads to a weaker relation between the latent trait and the item response; therefore, the item discrimination will be lower. Because the Bradlow et al. model contradicts the empirical results, it will not be considered further.

sponses, $E[\text{cor}(X_1,X_2)]$, is computed for the case in which the subjects are normally distributed. Second, the expected value of a correlation between two item responses after conditioning on the latent trait, $E[\text{cor}(X_1,X_2|\theta)]$, is calculated. After the common influence of the latent trait is removed, and if the two items do not interact, $E[\text{cor}(X_1,X_2|\theta)] = 0$. However, if some interaction remains, $E[\text{cor}(X_1,X_2|\theta)]$ may differ from 0.

The expected values can be found in Table 1. For simplicity, it is assumed that $\beta_1 = \beta_2 = 0$ and $\alpha_1 = \alpha_2 = 1$. The second column of Table 1 shows that a negative β_{12} results in a positive interaction and, hence, a larger correlation between the items than for $\beta_{12} = 0$ (the CI case); a positive β_{12} has the reverse effect. In the third column, it can be seen that if $\beta_{12} \neq 0$, there always remains some nonnegligible correlation between the two item responses. Hence, the latent trait can only account for all covariation between the item responses when $\beta_{12} = 0$.

Mathematical Analysis

In this section, a mathematical derivation is presented to show the extent to which item interactions can influence the magnitude of the estimated discrimination parameter. Readers who wish to see only the results of the analysis may skip (without loss of continuity) to the concluding paragraph of this section.

Take the constant interaction model for two items as a starting point. The marginal probability of a correct answer on Item 1 can be written as follows:

$$\begin{aligned}\Pr(X_1 = 1|\theta) &= \Pr(X_1 = 1, X_2 = 0|\theta) + \Pr(X_1 = 1, X_2 = 1|\theta) \\ &= \frac{\exp(\alpha_1(\theta - \beta_1)) + \exp((\alpha_1 + \alpha_2)\theta - \alpha_1\beta_1 - \alpha_2\beta_2 - \beta_{12})}{1 + \sum_{j=1}^{2}\exp(\alpha_j(\theta - \beta_j)) + \exp(\alpha_1(\theta - \beta_1) + \alpha_2(\theta - \beta_2) - \beta_{12})}.\end{aligned} \tag{4}$$

Table 1
Expected Correlations Between Item Responses Without and With Conditioning on the Latent Trait

β_{12}	$E[\text{cor}(X_1,X_2)]$	$E[\text{cor}(X_1,X_2|\theta)]$
−2	.544	.345
−1	.387	.197
0	.174	.000
1	−.062	−.197
2	−.257	−.345

An item that interacts with another item, but which is nevertheless analyzed as a separate item, is called a *marginal (interaction) item* in this article. Only the first item is taken into account in this derivation, but the line of argument for the second item is analogous (given the necessary changes of indexes) because the model in Equation 3 is symmetrical with respect to the items. To simplify the calculations, the following reparameterization is used:

$$\begin{aligned}e_1(\theta) &= \exp(\alpha_1(\theta - \beta_1)) \\ e_2(\theta) &= \exp(\alpha_2(\theta - \beta_2)) \\ e_{12} &= \exp(-\beta_{12}).\end{aligned} \tag{5}$$

Substituting these terms into Equation 4, we obtain the following simplified expression:

$$\Pr(X_1 = 1|\theta) = \frac{e_1(\theta) + e_1(\theta)e_2(\theta)e_{12}}{1 + e_1(\theta) + e_2(\theta) + e_1(\theta)e_2(\theta)e_{12}}. \tag{6}$$

For a binary 2PLM item (Equation 1) with difficulty parameter β_v and discrimination parameter α_v, the following holds for the log odds ratio of a correct versus incorrect answer:

$$\ln\left(\frac{\Pr(X_v = 1|\theta)}{\Pr(X_v = 0|\theta)}\right) = G(\theta) = \alpha_v(\theta - \beta_v). \tag{7}$$

Next, we take the first derivative of Equation 7 to find the slope:

$$\frac{dG(\theta)}{d\theta} = \alpha_v. \tag{8}$$

The log odds are a linear function of θ with slope α_v and intercept $-\alpha_v\beta_v$.

Similarly, the log odds for the first marginal item can be calculated with the probability formula given in Equation 6. The first derivative with respect to θ of the log odds function reveals something about the way the interaction influences the estimated discrimination parameter of the marginal item. The log odds function for the first marginal item is as follows:

$$\begin{aligned}\ln\left(\frac{\Pr(X_1 = 1|\theta)}{\Pr(X_1 = 0|\theta)}\right) &= F(\theta) \\ &= \ln\left(\frac{e_1(\theta) + e_1(\theta)e_2(\theta)e_{12}}{1 + e_2(\theta)}\right).\end{aligned} \tag{9}$$

Figure 1 shows the log odds function $F(\theta)$ for three marginal items with different values for the interaction parameter. In all three cases, the difficulty parameter of the item equaled 0 and the true discrimination parameter equaled 1. The marginal item with

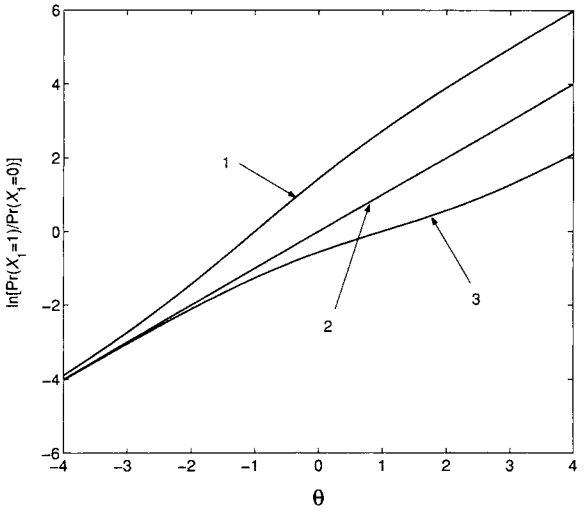


Figure 1. The log odds for a correct versus an incorrect response on a marginal item for three different values of the interaction parameter: Line 1 shows a positive interaction, $\beta_{12} = -2$; Line 2 shows a two-parameter logistic model item, $\beta_{12} = 0$; and Line 3 shows a negative interaction, $\beta_{12} = 2$. The other parameters are the same for the three curves: $\alpha_1 = \alpha_2 = 1$ and $\beta_1 = \beta_2 = 0$.

$\beta_{12} = 0$ is equal to a 2PLM item, and it is shown as a basis for comparison.

The first derivative of the log odds function $F(\theta)$ with respect to θ equals:

$$\frac{dF(\theta)}{d\theta} = \alpha_1 + \frac{\alpha_2 e_2(\theta)(e_{12} - 1)}{(1 + e_2(\theta))(1 + e_2(\theta)e_{12})}. \quad (10)$$

Hence, the slope depends on $e_2(\theta)$ but not on $e_1(\theta)$; therefore, it depends on the difficulty of the second item but not on that of the first item. It can be shown that the slope of the log odds function $F(\theta)$ for any θ is bounded as follows:

$$\alpha_1 < dF(\theta) / d\theta < U \quad \text{if } \beta_{12} < 0$$

$$L < dF(\theta) / d\theta < \alpha_1 \quad \text{if } \beta_{12} > 0, \quad (11)$$

where U is a finite upper bound and L is a lower bound larger than zero. From Equation 10, the bounding by α_1 in both cases may easily be checked (the sign of the second term in the right-hand side depends on the difference between $e_{12} - 1$). If $\beta_{12} = 0$, the first derivative is always equal to α_1. The full proof of Equation 11 can be found in Appendix A.

In general, if the log odds function $F(\theta)$ for the first marginal item is approximated by a linear log odds function $\lambda\theta + \mu$ (hence, a 2PLM item with discrimination λ and difficulty $-\mu/\lambda$), the following relations between λ and α_1 hold:

$$\begin{cases} \text{if } \beta_{12} < 0 \text{ then } \lambda > \alpha_1 \\ \text{if } \beta_{12} = 0 \text{ then } \lambda = \alpha_1 \\ \text{if } \beta_{12} > 0 \text{ then } \lambda < \alpha_1. \end{cases} \quad (12)$$

On one hand, a positive interaction will result in an overestimation of the discrimination parameter of an item for which the interaction is ignored. On the other hand, a negative interaction will result in an underestimation of the discrimination parameter. The relations in Equation 12 are derived in Appendix B. Because it is difficult to determine an exact value of λ given β_{12} with a mathematical analysis (because of the influence of many variables such as β_2, distribution of θ values, etc.), a numerical technique and a simulation study are performed in the next two sections.

In conclusion, there are two important results from our mathematical analysis: First, the item discrimination depends on the interaction parameter β_{12}. This was an expected finding, which has already been described in the literature, although it was not accomplished by mathematical analysis. This result can also be made clearer in an intuitive way. If two items interact positively, they provide less information than do two independent items. Hence, when they are analyzed as if they were independent, the information they supply about the latent trait (as measured by their item discrimination) is overestimated. If the two items interact negatively, the information they provide when analyzed separately is underestimated. This is a bit more difficult to understand than the case of positively interacting items. Assume for simplicity that all items are of equal difficulty. A negative interaction introduces additional uncertainty in the item responses. A person with a large θ may respond correctly on the first item but incorrectly on the second one of equal difficulty, because of the negative interaction, whereas it would be expected, on the basis of ability, that both responses would be correct. For a person with a small θ, an incorrect response may be followed by a correct one, despite the small latent scale value. Because of this additional uncertainty in the item responses, analyzing the items separately as if they were independent implies an underestimation of the information the items provide about the latent trait. The mathematical analysis presented above confirms this logic.

Second, the item discrimination parameter of an

interacting item also depends on the difficulty of the item it interacts with, but not on its own difficulty (see Equation 10). This relation is further investigated in the following sections. This effect of item difficulty on the estimated discrimination parameter in the case of ignored item interactions can be deduced from evidence in the literature. It has been noted by Douglas, Kim, Habing, and Gao (1998) that item difficulty has an influence on the magnitude of the LID between two items and, therefore, an effect on the estimated discrimination parameter. However, it is surprising that only the difficulty of the other item in the interaction has an influence on the estimated discrimination parameter.

Comparison of the ICCs With Curve Fitting

We now take the analysis one step further and argue that the ICCs of a 2PLM item and a marginal interaction item (involved in an interaction with another item) can be very similar. To this end, we approximate the ICC of a marginal item with the ICC of a 2PLM item by using a numerical optimization procedure. This is done for different sets of parameter values, and a 2PLM ICC often shows a good approximation to the ICC of the marginal item.

For every pair of interacting items, there are two marginal interaction items, and both can be approximated by a separate 2PLM item. Suppose Items 1 and 2 interact with each other according to the model in Equation 3, so that there are five item parameters in the model: α_1, β_1, α_2, β_2, and β_{12}. Suppose further that the interaction is ignored and that one tries to approximate the two marginal interaction items by using two 2PLM items. The parameters that yield the best approximation to the ICCs of the marginal items are denoted as values $\hat{\alpha}_1$, $\hat{\beta}_1$, $\hat{\alpha}_2$, and $\hat{\beta}_2$.

To show the degree of approximation between the ICCs of the marginal and the approximating 2PLM items, a distance function that measures the deviation between these ICCs is minimized. A general description of the minimization procedure can be found in Verhelst, Glas, and de Vries (1997). However, our approach (Tuerlinckx, 2000) was slightly different, thus the values of the minimized deviance reported in this article are not directly comparable to the values from Verhelst et al.

The minimization of the deviation between the ICCs is accomplished for different parameter values of the interaction model. The values for the discrimination parameters in the interaction model can be 0.5, 0.8, 1.0, 1.2, or 1.5 (given a standard normal distribution for θ, these values range from weakly to strongly discriminating), and the difficulty and interaction parameters can be −2.0, −1.0, 0.0, 1.0, or 2.0. There are five different parameters (α_1, β_1, α_2, β_2, and β_{12}), and for each parameter there are five possible parameter values, yielding a total of $5 \times 5 \times 5 \times 5 \times 5 = 3{,}125$ different combinations. For each combination, the minimization routine is applied. The deviance value between the ICCs is denoted by F^{tot} (F^{tot} is based on the integrated squared distance between two ICCs weighted with the normal density).

To illustrate our approach, we show the ICCs of three pairs of marginal interaction items together with the approximated 2PLM items in Figure 2. The first pair is chosen for its rather low discrepancy, the second for its average discrepancy (close to the overall mean of 0.017), and the third for its rather high discrepancy.

The first pair of items comes from an interaction model with parameters $\alpha_1 = 0.5$, $\beta_1 = 0.0$, $\alpha_2 = 0.5$, $\beta_2 = 0.0$, and $\beta_{12} = -2.0$. Panel A contains the trace line for the first approximating 2PLM item, with an estimated discrimination of 0.72 and a difficulty of 0.00. Panel B shows the second approximating 2PLM item, with the same estimated parameters as the first

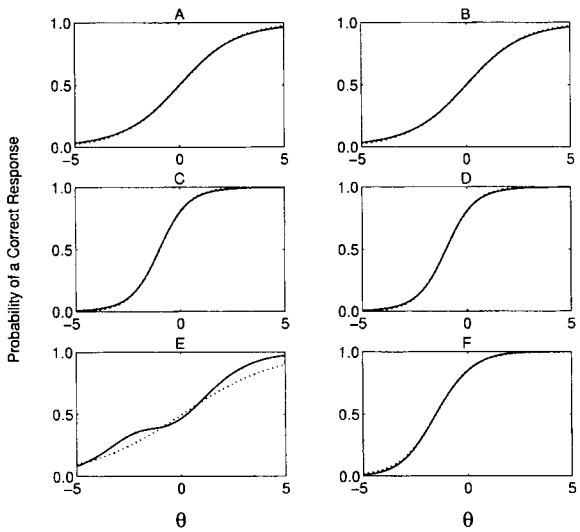


Figure 2. The item characterstic curves for a two-parameter logistic model item (dashed line) and a marginal interaction item (solid line). Panels A and B show an item pair with a good match, $F^{tot} = 0.004$; Panels C and D show an item pair with a moderate match, $F^{tot} = 0.015$; and Panels E and F show a case with a poor match, $F^{tot} = 0.066$.

one. The close match between the curves for this first pair indicates that the 2PLM items provide fairly close approximations to the marginal interacting items.

The true parameter values for the second pair of interacting items are $\alpha_1 = 1.0$, $\beta_1 = 0.0$, $\alpha_2 = 1.0$, $\beta_2 = 0.0$, and $\beta_{12} = -2.0$. In Panel C, the approximating 2PLM curve to the first item of the pair is shown; it has estimated parameter values of $\hat{\alpha}_1 = 1.41$ and $\hat{\beta}_1 = -1.00$. Panel D shows the approximating 2PLM item for the second marginal interaction item. The estimated 2PLM parameter values for the second item are exactly the same as for the first one. Hence, the match between the curves for the second pair, which has a much higher F^{tot}, is also quite good.

For the third pair, the true parameter values in the interaction model are $\alpha_1 = 0.8$, $\beta_1 = -2.0$, $\alpha_2 = 1.5$, $\beta_2 = -2.0$, and $\beta_{12} = 2.0$. The curves for the first item are given in Panel E. The marginal item from an interaction model is not closely approximated by a 2PLM item (estimated parameter values are $\hat{\alpha}_1 = 0.44$ and $\hat{\beta}_1 = 0.05$). Panel F shows the second marginal interaction item and the approximating 2PLM item (estimated parameter values are $\hat{\alpha}_1 = 1.15$ and $\hat{\beta}_1 = -1.51$). Mainly because of poor approximation of the first marginal interacting item, the total mismatch is considerably larger for the third pair, as reflected in $F^{tot} = 0.066$.

In Table 2, the effects on $\hat{\alpha}_1$, $\hat{\beta}_1$, and F^{tot} are shown for the different values of α_1, β_1, and β_{12}. The effects of α_2, β_2, and β_{12} on the estimated parameters of the second item are not shown, but because the problem is symmetrical, the results are analogous.

Several conclusions can be drawn from Table 2. First, the two factors that most influence the estimated discrimination parameter of the first item ($\hat{\alpha}_1$) are the true discrimination parameter of the item, α_1, and the interaction parameter, β_{12}. If the interaction becomes more positive (β_{12} decreases), the estimated discrimination parameter increases. This was already concluded from our theoretical analysis in the previous section. The difficulty parameter β_1 has no influence on $\hat{\alpha}_1$. Although β_2 appears in Equation 10, there is no main effect of β_2 on α_1 (not shown in Table 2); this will be clarified in the simulation study.

Second, the two factors that most influence the estimated difficulty of the first item ($\hat{\beta}_1$) are the true difficulty parameter of the item, β_1, and the interaction parameter, β_{12}. The main effect of β_{12} on $\hat{\beta}_1$ is approximately $\beta_{12}/2$. No systematic effects from the other manipulations were found. Third, there are no large main effects of the parameters of one item alone on the mean of the discrepancy value F^{tot}. Only the interaction parameter has a clear systematic effect on F^{tot}: The more extreme the interaction parameter, the larger the minimized value of the loss function. Moreover, the discrepancy value is larger for positive values of β_{12} than for negative values.

Table 2
Influence of the Parameters of an Interaction Model on the Estimated Discrimination and Difficulty Parameters of the 2PLM for the First Item and on the Value of the Loss Function (F^{tot})

Effect	Mean α_1	Mean $\hat{\beta}_1$	F^{tot}	N^a
α_1				
0.5	0.55	−0.19	0.015	558
0.8	0.82	−0.05	0.016	601
1.0	1.02	0.01	0.017	610
1.2	1.22	−0.01	0.019	608
1.5	1.52	−0.02	0.021	602
β_1				
−2.0	1.03	−2.04	0.016	596
−1.0	1.03	−1.04	0.018	593
0.0	1.03	−0.03	0.018	595
1.0	1.03	0.98	0.019	606
2.0	1.05	1.90	0.016	589
β_{12}				
−2.0	1.31	−1.10	0.022	624
−1.0	1.16	−0.56	0.015	625
0.0	1.00	0.00	0.000	625
1.0	0.86	0.60	0.019	615
2.0	0.77	1.09	0.037	490

[a] This column shows the number of converged solutions (out of 625). The technique sometimes fails to render realistic values for the two 2-parameter logistic model (2PLM) items, especially for positive values of the interaction parameter; thus those outcomes are discarded.

In sum, there exist several combinations of parameter values in the interaction model that allow the ICCs of the marginal items to be almost perfectly mimicked by the 2PLM, but for other combinations the discrepancy between the models is much larger.

Simulation Studies

In the previous sections, the interacting items were treated in isolation, and the analyses were mainly theoretical. The question could be raised to what extent the phenomenon under study shows up in practice when the interacting items are embedded in a set of other noninteracting items, or when more pairs of interacting items are present. These questions are addressed in three simulation studies.

There are two dependent variables of interest in the simulations. First, we compare the estimated value of the discrimination parameter of the items involved in an interaction to their true value. Second, we evaluate the rejection rate of the likelihood ratio test for the goodness-of-fit of the 2PLM model.

In this simulation study, we assume that the true item discrimination is equal to 1 for all (interacting and noninteracting) items. This is a very interesting case, in which the only reason for differences in the estimated discriminations across items is the ignored interaction.

Simulation Study 1

The first simulation study had three objectives. First, in the previous analyses, it was shown that the magnitude of the interaction has a systematic effect on the estimated item discrimination. Thus, we wanted to discover whether this effect can be replicated with simulated data. Second, it was also concluded that the difficulty of the other item in the interaction has an influence on the estimated item discrimination; this also requires confirmation in simulated data. Third, in an exploratory fashion, we investigated the influence of the number of subjects on the estimated item discrimination and rejection rates.

Data. Binary data were generated for six items, with the restriction that all item discriminations were equal to 1. Hence, in this study, the effects on the estimated item discrimination can be directly compared to a common baseline. An interaction model such as the one in Equation 3 was assumed for Items 1 and 2 ($\alpha_1 = \alpha_2 = 1$). The model for the four items not involved in any interaction was a CI model such as the one in Equation 1 (with $\alpha_i = 1$). Thus, there were two subsets: one containing the two interacting items (testlet) and the other containing the four remaining CI items.

For each separate cell in the simulation studies, 50 replications were generated and analyzed. Because the interaction model is symmetrical for two interacting items, we look only at the effects of the interaction on the estimated item discrimination of Item 1.

There were three independent variables in the first simulation study (listed together with their values): the interaction parameter ($\beta_{12} = -2, -1, 0, 1, 2$), the item parameter of the second item ($\beta_2 = -2, -1, 0, 1, 2$), and the number of subjects ($n = 500$ or 1,000).

The difficulty parameter of Item 2, β_2, was manipulated to explore its effect on the estimated discrimination parameter of Item 1. The three independent variables were combined in a factorial design. The difficulty parameters of Item 1, β_1, and of Items 3–6 were randomly drawn from a standard normal distribution. The randomly drawn values varied across replications. The latent trait values θ were also drawn from a standard normal density.

Method. The 2PLM was fitted by means of MULTILOG (Thissen, 1991), which yields (marginal) maximum likelihood estimates for the item parameters of the 2PLM. The output from MULTILOG also contains the likelihood ratio test statistic:

$$G^2 = 2\sum_{\boldsymbol{x}} n_{\boldsymbol{x}} \ln\left(\frac{n_{\boldsymbol{x}}}{\hat{n}_{\boldsymbol{x}}}\right), \tag{13}$$

where $n_{\boldsymbol{x}}$ is the observed frequency of the response vector $\boldsymbol{x}$, and $\hat{n}_{\boldsymbol{x}}$ is the expected frequency under the model (the sum is taken over all possible response patterns $\boldsymbol{x}$). Under the null hypothesis that the model is true, G^2 is asymptotically chi-square distributed, with degrees of freedom equal to the number of free response patterns minus the number of free parameters in the model.

First, for each simulated data set, we compared the estimated discrimination parameter of Item 1, which is involved in an interaction, to the mean of the estimated discrimination parameters of the CI items. We expected estimated item discriminations of the noninteracting items to be close to their true value (i.e., 1), whereas the estimated item discriminations of the interacting Item 1 would be affected by the interaction.

Second, the rejection rates of the likelihood ratio test were investigated. More specifically, for each cell in the design (i.e., using the 50 replications), we calculated the proportion of times the observed value of the test statistic G^2 was larger than the critical value at $\alpha = .05$. This is an estimate of the true probability of exceeding the critical value. Hence, if the model holds, we expect about 5% exceedings. However, two remarks are necessary. First, it should be kept in mind that the assumptions of the likelihood ratio test are probably not met because the sample size was finite and far from large. Hence, the chi-square approximation may not be accurate, and, as a result, the estimated proportion of exceedings may deviate from its true value. Secondly, only 50 replications per cell were computed, which may be too few to obtain stable estimates of the rejection rates (particularly if these are close to 0 or 1). However, our intention was

Table 3
Mean (± SD) Estimated Discrimination Parameters for Every Combination of β_{12} and β_2 for a Marginal Interacting Item and for the Noninteracting Items (n = 500)

	β_2				
β_{12}	−2	−1	0	1	2
−2	1.96 ± 0.61	2.79 ± 1.28	3.78 ± 2.06	3.77 ± 1.68	3.00 ± 0.67
	0.92 ± 0.19	0.82 ± 0.18	0.79 ± 0.16	0.77 ± 0.18	0.77 ± 0.17
−1	1.32 ± 0.28	1.66 ± 0.38	1.79 ± 0.39	1.72 ± 0.35	1.55 ± 0.32
	0.95 ± 0.22	0.91 ± 0.20	0.89 ± 0.17	0.90 ± 0.18	0.94 ± 0.19
0	1.03 ± 0.22	1.05 ± 0.24	0.96 ± 0.25	1.03 ± 0.33	1.00 ± 0.23
	1.02 ± 0.24	1.03 ± 0.22	1.02 ± 0.22	1.01 ± 0.23	1.00 ± 0.23
1	0.77 ± 0.20	0.63 ± 0.21	0.63 ± 0.19	0.73 ± 0.19	0.87 ± 0.22
	1.06 ± 0.27	1.06 ± 0.22	1.06 ± 0.26	1.05 ± 0.27	1.03 ± 0.25
2	0.47 ± 0.25	0.32 ± 0.21	0.43 ± 0.28	0.60 ± 0.21	0.81 ± 0.23
	1.06 ± 0.32	1.06 ± 0.30	1.05 ± 0.25	1.06 ± 0.25	1.07 ± 0.40

Note. The top value in each cell is the mean (± *SD*) of $\hat{\alpha}_1$. The bottom value is the mean (± *SD*) of the estimated item discrimination parameters of Items 3–6.

simply to gain a tentative idea of the rejection rates and changes in rejection rates over conditions.

Results. The mean item discriminations for the first marginal item and for the CI items (Items 3–6) can be found in Table 3. Because the number of subjects did not affect the estimated discrimination parameter (there was only an obvious effect on the standard deviation), only the results for $n = 500$ are given.

It can be seen in Table 3 that the value of the interaction has the expected effect on the estimated discrimination parameter for the first item: The estimated item discrimination is larger than 1 when the interaction is positive ($\beta_{12} < 0$), it is smaller than 1 when the interaction is negative ($\beta_{12} > 0$), and there is no effect when CI holds among all items. This confirms the results of the previous theoretical analyses. However, there is also a small effect of the interaction parameter on the mean item discrimination of the CI items. A trade-off relation is induced between the estimated discrimination parameters of the interacting items and noninteracting items because a restriction must be imposed on the item discriminations to fix the latent scale of the model; this is necessary to obtain identifiability.

The effect of the interaction parameter on the estimated discrimination of the first item is moderated by the difficulty of the second item. In general, a more extreme value of the difficulty parameter reduces the effect of the interaction. This can also be deduced from Equation 10. If β_2 becomes very large or very small (by taking the limit to ∞ and $-\infty$, respectively), $dF(\theta)/d\theta$ approaches α_1, which shows that extreme values of β_2 reduce the effect of the interaction.

The proportion of cases in which G^2 exceeds the critical value at $\alpha = .05$ is tabulated for each condition in Table 4 for $n = 500$. The $\beta_{12} = 0$ row gives the estimated probability of a Type I error given the nominal significance level of $\alpha = .05$. Except for one accidental case,[3] the estimated probability is close to .05. The rejection rates in the other rows refer to the power of the test because those are the proportions of cases in which the wrong model is falsified. It can be seen that there are many more significant cases for positive values of the interaction parameter (i.e., negative interactions) than for negative values (i.e., positive interactions). This confirms a result also shown in Table 2, in which negative interactions ($\beta_{12} > 0$) give rise to somewhat larger discrepancies between the two types of ICCs.

For $\beta_{12} = 0$, the rejection rates for $n = 1{,}000$ are the same magnitude as those for $n = 500$ (i.e., when the correct model is fitted), but they are larger for the wrong models ($\beta_{12} \neq 0$). In general, the same qualitative patterns for the rejection rates hold as for $n = 500$.

Simulation Study 2

In the second simulation study, the effect of the number of items was investigated. Therefore, the

[3] As a check on this anomalous result, we performed a second set of 250 simulations for the cell ($\beta_{12} = 0, \beta_2 = 0$) in Table 4, in which only 10 out of 250 likelihood ratio tests (4%) were significant at $\alpha = .05$.

Table 4
Proportion of Significant Likelihood Ratio Tests at α = .05 for Different Values of β_{12} and β_2 (Six Items, n = 500)

	β_2				
β_{12}	−2	−1	0	1	2
−2	.18*	.22*	.30*	.38*	.48*
−1	.14*	.22*	.26*	.22*	.14*
0	.02	.00	.16*	.02	.08
1	.24*	.52*	.30*	.28*	.08
2	.78*	.88*	.94*	.78*	.28*

Note. Asterisks indicate that the reported proportion of significant results is significantly different from .05 at α = .05.

number of CI items was manipulated, but the number of interacting items remained at two (hence the subset of CI items is made larger).

Data and method. Data generation was largely the same as in the first simulation study. The only difference was that the number of items was now manipulated: There were either 10 or 20 items. As in the first simulation study, there was an interaction between Items 1 and 2 and the other items were CI items. In addition, the magnitude of the interaction between Items 1 and 2 was manipulated (β_{12} = −2, −1, 1, or 2). Because β_{12} = 0 corresponds to a CI model for all items, this value was not considered. The difficulties of all individual items (β_is) were randomly drawn from a normal distribution (hence they are not manipulated). The method of analysis remained the same as in the first study.

Results. As can be seen in Table 5, the effect of the interaction of the items on the estimation of the discrimination parameter is reduced if the number of items increases. The second column contains the results from the first simulation study, with 6 items and n = 500 (see Table 3, but results averaged over β_2). The estimated item discrimination for Item 1 deviates more from its true value in a smaller set of items. This may be explained as follows: The slope of the log odds function in Equation 10 depends on θ. However, the estimate of θ will be biased, as the wrong model (2PLM) is assumed. The relative size of the bias will decrease as the item set gets larger because the relative contribution of the interacting items that cause the bias decreases as the number of CI items increases. Therefore, the estimated discrimination parameter will be less biased if more CI items are considered. Nevertheless, even with 20 items, the interaction has large systematic effects on the estimated discrimination parameter.

Table 5
Mean (± SD) Estimated Discrimination Parameters for k = 6, k = 10, and k = 20 Items for 500 Examinees (Effects on Interacting Items and on Other Items)

	Number of items		
β_{12}	6	10	20
−2	3.06 ± 1.26	2.28 ± 0.50	1.63 ± 0.25
	0.82 ± 0.18	0.93 ± 0.12	0.99 ± 0.11
−1	1.61 ± 0.34	1.42 ± 0.24	1.26 ± 0.13
	0.93 ± 0.28	0.98 ± 0.13	1.00 ± 0.10
1	0.72 ± 0.20	0.72 ± 0.13	0.77 ± 0.12
	1.05 ± 0.26	1.02 ± 0.14	1.00 ± 0.11
2	0.53 ± 0.24	0.57 ± 0.14	0.59 ± 0.14
	1.06 ± 0.30	1.02 ± 0.13	1.00 ± 0.11

Note. The top value in each cell is the mean (± *SD*) of $\hat{\alpha}_1$. The bottom value is the mean (± *SD*) of the estimated discrimination parameters for Items 3–6 (if k = 6), 3–10 (if k = 10), or 3–20 (if k = 20).

No likelihood ratio test was carried out in this case because the contingency table for testing the model would contain too many cells compared to the number of examinees, so that the asymptotic approximation would probably no longer be valid (e.g. for k = 20, there are 2^{20} different response patterns for only 500 examinees). Given that it is difficult to test the model if the number of items is 10 or more, the effect of ignored interactions will go mostly undetected in those cases.

Simulation Study 3

In the previous analyses, only one pair of interacting items was studied. The objective of the third simulation study was to investigate the combined effect of multiple two-item interactions. The total item set was divided into three subsets: The first subset contains the CI items, the second subset contains the first pair of interacting items, and the third subset contains the second pair of interacting items.

Data and method. As in the previous studies, data for six items were generated, and interactions between Items 1 and 2 and between Items 3 and 4 were assumed. For both interactions, the interaction parameters (β_{12} and β_{34}) were each independently varied in a factorial design with possible values of −2, −1, 1, or 2. There were 500 examinees. The individual difficulty parameters were randomly drawn from a normal distribution. The method of analysis was the same as in the first study.

Results. The results on the estimated discrimination parameters are not discussed in detail because they are similar to what was found previously: A positive interaction results in an overestimation of the discrimination parameter and a negative interaction in an underestimation. The effect remains largely the same for one or two 2-item interactions.

Table 6 contains the number of significant likelihood ratio tests at level $\alpha = .05$ for two 2-item interactions. Compared with the case of one 2-item interaction, the rejection rate is higher, so that one may conclude that the rejection rate increases with the number of interactions. Table 6 shows that combinations of large absolute values of β_{12} and β_{34} also result in larger rejection rates.

Conclusion

The systematic effect of ignoring item interactions on the estimated discrimination parameter in the 2PLM that was found in the theoretical studies was confirmed in the simulation study. Moreover, the simulation study shows that the same effects emerge when more than one 2-item interaction is present in the data set. When the interaction is embedded in a larger set of CI items, the effect on the item discrimination remains, but it is less pronounced.

In addition, the power of the likelihood ratio test with respect to the presence of interactions in the data was explored. In general, more interactions and more extreme interactions are more often detected as violations of the 2PLM, and negative interactions lead to higher rejection rates than do positive interactions. However, in many cases the item interactions remain undetected.

Discussion

The theoretical analysis and the simulation study clearly show that neglecting interactions between items may result in false conclusions about the discrimination parameters of the items in a given test. Positive interactions cause the discrimination parameters to be overestimated, and negative interactions yield underestimated discrimination parameters. Furthermore, the 2PLM can often approximate the underlying interactions quite closely, so that a goodness-of-fit test of the 2PLM cannot detect the discrepancy with the true model.

Table 6
Proportion of Significant Likelihood Ratio Tests at $\alpha = .05$ for Two 2-Item Interactions (Six Items, 500 Examinees)

	β_{34}			
β_{12}	−2	−1	1	2
−2	.92	.88	.60	.90
−1	.80	.58	.32	.90
1	.40	.38	.66	.96
2	.74	.92	.96	1.00

Note. All proportions differ significantly from .05 at $\alpha = .05$.

Interaction effects are likely to occur in a number of settings. First, they may occur if a set of items has a common stem, as in the case of questions that follow the same reading passage or picture. More than one question is asked about the reading passage because of economic and time constraints, but, if two or more questions address the same topic, an interaction may result. Second, interactions may originate from subtle carryover effects caused by the specific order of test taking. The correct response on one question may contain elements that are helpful in correctly answering the next one.

Two caveats should be added to these conclusions to avoid misunderstanding. First, it should not be concluded from our results that differences in discrimination across items are always attributable to interactions. We do not deny that items can have genuine variation in their discrimination parameters. Second, neglecting interactions between items can also affect other quantifications of test data that were not investigated here (or only superficially) but were reported in the literature overview, such as information about estimated difficulty parameters, θ, reliability measures, and so on. For instance, the item information $J_i(\theta)$ for item i in the 2PLM is defined as

$$J_i(\theta) = \alpha_i^2 \Pr(X_i = 1|\theta)(1 - \Pr(X_i = 1|\theta)). \quad (14)$$

It can be algebraically shown that if α_i is increased (decreased), the information for the θ values in the neighborhood of the item difficulty increases (decreases) also. This shows that overestimating (underestimating) α_i also causes inflation (deflation) of the item information for some range of θ values. As a consequence, the test information

$$\sum_i J_i(\theta) \quad (15)$$

for the θ values in the difficulty range is also inflated (deflated) and the standard errors of the estimated θ values become smaller (larger). In addition, the estimated difficulty parameter (as shown in the curve approximation section) can be influenced by the in-

teraction. The same holds true for the estimated person parameter.

An important reason why we chose item discrimination as the variable of interest in this study is that it provides information about the quality of the item. Items with large discriminations are widely considered to be better than those with smaller discriminations because they provide a purer measure of the latent trait. However, it is shown here that a high discrimination is not necessarily good. The discrimination parameter is shown to vary systematically with the interaction between items, so that if there is interaction, the discrimination can be seriously biased; this bias is especially liable to go undetected when one is not looking for interactions. When a high discrimination stems from interaction, then it is mainly an indication of redundancy with one or more other items. The results reported in this article may be of particular interest to users of item banks. Suppose that two interacting items are included in the item bank and, in a later stage, only one of them is chosen to be presented to a test taker. The new properties of this item that is isolated from the other item of the interaction pair may be unpredictable. The same problem arises with computerized adaptive testing, in which the wrong items may be chosen because of a biased item-information curve.

A researcher who wants to apply the 2PLM to test data has to keep in mind the possibility of interactions (or more broadly, LIDs) and their distorting influence on the parameter estimates (certainly those related to the quality of the item). It is often very difficult to detect the presence of item interactions before the test is taken, but there are situations where one might expect item interactions; a few of them are mentioned in the introduction and previously in the Discussion. The first step in the detection process is often to apply an omnibus test such as the likelihood ratio test (Equation 13). Although it is shown in this article that an omnibus test may give misleading results in some cases, there are other cases in which the test may indeed reveal a problem. Besides such omnibus tests, more specific tests are also available for the detection of LIDs (Chen & Thissen, 1997; Yen, 1984), and we believe it is a good practice to also check the results of the IRT analyses with these methods precisely because they are more specific and more powerful. As a last step, the interaction model presented in this article could be fitted to the data, and the inspection of the interaction parameter values should give an indication of the presence of interactions in the data. Hoskens and De Boeck (1997) discuss how these models can be estimated and tested.

References

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–424). Reading, PA: Addison-Wesley.

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

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265–289.

Douglas, J., Kim, H. R., Habing, B., & Gao, F. (1998). Investigating local dependence with conditional covariance functions. *Journal of Educational and Behavioral Statistics, 23*, 129–151.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Holland, P., & Rosenbaum, P. (1986). Conditional association and unidimensionality in monotone latent variable models. *Annals of Statistics, 14*, 1523–1543.

Hoskens, M., & De Boeck, P. (1995). Componential IRT models for polytomous items. *Journal of Educational Measurement, 32*, 234–246.

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

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

Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika, 49*, 223–245.

Knowles, E. S., & Condon, C. A. (2000). Does the rose still smell as sweet? Item variability across test forms and revisions. *Psychological Assessment, 12*, 245–252.

Rosenbaum, P. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika, 49*, 425–435.

Schwarz, N. (1999). How the questions shape the answers. *American Psychologist, 54*, 93–105.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet based tests. *Journal of Educational Measurement, 28*, 237–247.

Steinberg, L., & Thissen, D. (1996). Uses of item response

theory and the testlet concept in the measurement of psychopathology. *Psychological Methods, 1*, 81–97.

Thissen, D. (1991). MULTILOG [Computer software]. Mooresville, IN: Scientific Software.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement, 26*, 247–260.

Tuerlinckx, F. (2000). *Some extensions and applications of item response theory models and diffusion models*. Unpublished doctoral dissertation, University of Leuven, Leuven, Belgium.

Verhelst, N., Glas, C. A. W., & de Vries, H. H. (1997). A steps model to analyze partial credit. In W. J. Van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 123–138). New York: Springer.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185–201.

Wilson, M., & Adams, R. J. (1995). Rasch models for item bundles. *Psychometrika, 60*, 181–198.

Yen, W. M. (1984). Effect of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125–145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187–213.

Appendix A

Proof of Equation 11

The proof of Equation 11 is tedious but straightforward. The outline is sketched for the case in which $\beta_{12} < 0$. The lower boundary is easy to check. The upper boundary is obtained by proving that $a(\theta) = dF(\theta)/d\theta$ has a unique maximum. For this purpose, $a(\theta)$ is differentiated with respect to θ, and the resulting expression is set equal to zero and solved for θ. One can always find a unique root of $da(\theta)/d\theta$, and this root corresponds to the only maximum of $a(\theta)$. That the root corresponds to a maximum (and not to a minimum or inflection point) can be derived from the following facts:

$$\lim_{\theta \to +\infty} a(\theta) = \lim_{\theta \to -\infty} a(\theta) = \alpha_1$$

and $a(\theta) > \alpha_1$ for any θ. This proves that the slope $dF(\theta)/d\theta$ is bounded for the case $\beta_{12} < 0$; the other case is analogous.

Appendix B

Derivation of Equation 12

The derivation of the second part in Equation 12 is trivial; therefore, we will derive the first inequality. The third derivation is analogous. When the log odds function $F(\theta)$ is approximated by a linear function $\lambda\theta + \mu$, the approximation error will be quantified by some distance function that measures the deviation of $F(\theta)$ to the approximation. This distance function may appear as follows:

$$H(\lambda,\mu) = \int_{-\infty}^{\infty} (F(\theta) - \lambda\theta - \mu)^2 \phi(\theta) d\theta, \qquad \text{(B1)}$$

where $\phi(\theta)$ is the weighting function (e.g., $\phi(\theta) = 1$ or ϕ is the normal density), $F(\theta)$ is the log odds function of the marginal item, and $\lambda\theta + \mu$ is the approximating line. We impose two conditions on the weight function:

$$\int_{-\infty}^{\infty} \theta\phi(\theta)\, d\theta = 0 \text{ and } \phi(\theta) = \phi(-\theta) \text{ (symmetry)}. \qquad \text{(B2)}$$

Furthermore, assume for simplicity that $\beta_{12} < 0$ and $\beta_1 = 0$.

Minimizing $H(\lambda,\mu)$ can be done independently for λ and μ; therefore, μ will not be considered anymore. After taking the first derivative of $H(\lambda,\mu)$ with respect to λ, setting it equal to zero and solving for λ, one obtains the following expression:

$$\lambda = \frac{\int_{-\infty}^{\infty} \theta F(\theta)\phi(\theta) d\theta}{\int_{-\infty}^{\infty} \theta^2 \phi(\theta) d\theta}. \qquad \text{(B3)}$$

The derivation is continued by inserting the formula for $F(\theta)$:

$$\lambda = \frac{\int_{-\infty}^{\infty}\theta\ln\left(e_1(\theta)+e_1(\theta)e_2(\theta)e_{12}\right)\phi(\theta)d\theta - \int_{-\infty}^{\infty}\theta\ln\left(1+e_2(\theta)\right)\phi(\theta)d\theta}{\int_{-\infty}^{\infty}\theta^2\phi(\theta)d\theta}$$

$$= \frac{\alpha_1\int_{-\infty}^{\infty}\theta^2\phi(\theta)d\theta + \int_{-\infty}^{\infty}\theta\ln\left(1+e_2(\theta)e_{12}\right)\phi(\theta)d\theta - \int_{-\infty}^{\infty}\theta\ln\left(1+e_2(\theta)\right)\phi(\theta)d\theta}{\int_{-\infty}^{\infty}\theta^2\phi(\theta)d\theta}$$

$$\Downarrow \quad \beta_{12} < 0 \Rightarrow e_{12} > 1$$

$$> \frac{\alpha_1\int_{-\infty}^{\infty}\theta^2\phi(\theta)d\theta + \int_{-\infty}^{\infty}\theta\ln\left(1+e_2(\theta)\right)\phi(\theta)d\theta - \int_{-\infty}^{\infty}\theta\ln\left(1+e_2(\theta)\right)\phi(\theta)d\theta}{\int_{-\infty}^{\infty}\theta^2\phi(\theta)d\theta} = \alpha_1. \quad \text{(B4)}$$

To justify this last step, it has to be shown that the following inequality holds:

$$\int_{-\infty}^{\infty}\theta\ln\left(1+e_2(\theta)e_{12}\right)\phi(\theta)d\theta > \int_{-\infty}^{\infty}\theta\ln\left(1+e_2(\theta)\right)\phi(\theta)d\theta \quad \text{(B5)}$$

or

$$\int_{-\infty}^{\infty}\theta[\ln\left(1+e_2(\theta)e_{12}\right)-\ln\left(1+e_2(\theta)\right)]\phi(\theta)d\theta = \int_{-\infty}^{\infty}\theta\ln\left(\frac{1+e_2(\theta)e_{12}}{1+e_2(\theta)}\right)\phi(\theta)d\theta > 0 \quad \text{(B6)}$$

if $\beta_{12} < 0$. This means that

$$\int_{0}^{\infty}\theta\ln\left(\frac{1+e_2(\theta)e_{12}}{1+e_2(\theta)}\right)\phi(\theta)d\theta > -\int_{-\infty}^{0}\theta\ln\left(\frac{1+e_2(\theta)e_{12}}{1+e_2(\theta)}\right)\phi(\theta)d\theta. \quad \text{(B7)}$$

A sufficient condition for this inequality to hold is that for any value of $\theta > 0$

$$\theta\ln\left(\frac{1+e_2(\theta)e_{12}}{1+e_2(\theta)}\right)\phi(\theta) > -(-\theta)\ln\left(\frac{1+e_2(-\theta)e_{12}}{1+e_2(-\theta)}\right)\phi(-\theta). \quad \text{(B8)}$$

After some algebra, this last inequality is shown to be equivalent with $\beta_{12} < 0$. This proves the first part of Equation 12. The derivation for $\beta_{12} > 0$ is analogous.

Received June 15, 2000
Revision received January 22, 2001
Accepted February 7, 2001 ■