PSYCHOMETRIKA—VOL. 73, NO. 4, 533–559
DECEMBER 2008
DOI: 10.1007/S11336-008-9092-X

RANDOM ITEM IRT MODELS

PAUL DE BOECK

K.U.LEUVEN

It is common practice in IRT to consider items as fixed and persons as random. Both, continuous and categorical person parameters are most often random variables, whereas for items only continuous parameters are used and they are commonly of the fixed type, although exceptions occur. It is shown in the present article that random item parameters make sense theoretically, and that in practice the random item approach is promising to handle several issues, such as the measurement of persons, the explanation of item difficulties, and trouble shooting with respect to DIF. In correspondence with these issues, three parts are included. All three rely on the Rasch model as the simplest model to study, and the same data set is used for all applications. First, it is shown that the Rasch model with fixed persons and random items is an interesting measurement model, both, in theory, and for its goodness of fit. Second, the linear logistic test model with an error term is introduced, so that the explanation of the item difficulties based on the item properties does not need to be perfect. Finally, two more models are presented: the random item profile model (RIP) and the random item mixture model (RIM). In the RIP, DIF is not considered a discrete phenomenon, and when a robust regression approach based on the RIP difficulties is applied, quite good DIF identification results are obtained. In the RIM, no prior anchor sets are defined, but instead a latent DIF class of items is used, so that posterior anchoring is realized (anchoring based on the item mixture). It is shown that both approaches are promising for the identification of DIF.

Key words: random effects, generalizability, measurement, LLTM, DIF.

The parameters in IRT models are of various kinds. The most common kinds are continuous, and either fixed or random. For example, both, abilities and difficulties are continuous in the common IRT application. Although most models work with continuous parameters, an important class of models contains also categorical parameters in the form of a categorical random variable for the latent class or mixture component membership. Well-known recent variants of the mixture model family are the cognitive diagnostic models (CDM) (Roussos, Templin & Henson, 2007). For both, continuous and categorical parameters, persons are mostly treated as random. Abilities are seen as random effects, and latent class membership is also a random variable. This is in contrast with how items are almost always treated as fixed. Random item parameters are uncommon. The purpose of this manuscript is to illustrate that random item parameters, continuous and categorical, can make sense, and can be a useful tool to solve problems that remained unsolved thus far. The illustrations are meant as preliminary explorations based on the data set used by De Boeck and Wilson (2004). Random item models are rather new territory in psychometrics and require further study. Here, it is shown they are promising.

1. Short History of Random Items

There is a little tradition growing with respect to random item parameters. The origin can be found in the need of psycholinguists to counter the "language-as-fixed-effect fallacy" (Clark, 1973).

Requests for reprints should be sent to Paul De Boeck, K.U.Leuven, Leuven, Belgium. E-mail: paul.deboeck@psy.kuleuven.be

> ... it does not take into account the fact that the items are sampled from a larger population of items. ... leads to sampling variance that must be taken into account. Otherwise, this variance will be confounded with the effect of the treatment variable. (Raaijmakers, Schrijnemakers, & Gremmen, 1999, p. 416)

The issue was first discussed by Coleman (1964). A recent publication that can be seen in this line is from Rouder, Lu, Speckman, Sun, Morey, and Naveh-Benjamin (2007).

In the psychometric literature, the issue of random items has been discussed by researchers interested in the change of abilities over time, using a longitudinal design (Albers, Does, Ombos, & Janssen, 1989). In such designs, one may not repeat the same items, and the alternative is to draw nonoverlapping random samples from the same population of items. The approach was described by Tan, Ambergen, Does, and Imbos (1999) as follows:

> It can be argued that the item difficulties can be considered as outcomes of a random variable with some underlying distribution, possibly the normal distribution. ... As a consequence the item parameters can be integrated out. (p. 211)

The main motive for the most recent appearance of random item approaches in the literature is handling item families. Item families are sets of items with sufficient communalities within the set and sufficient differentiation from other sets, in order to (1) consider the items from the same family as being sampled from the same population, a different one depending on the family; (2) concentrate on the family parameters, which are distribution parameters (family mean and variance), instead of concentrating on item-specific parameters. The item family concept is often based on the principle of item generation (Bejar, 1993; Embretson, 1999), which has been called also "item cloning" (Glas & van der Linden, 2003) and "AIG—automatic item generation" (Sinharay, Johnson, & Williamson, 2003), but it can also rely on the notion of an achievement target being expressed in a set of items that represents the target (Janssen, Tuerlinckx, Meulders, & De Boeck, 2000). An interesting asset is that one can generalize to other items from the same family. Most studies share that a Bayesian estimation method is used, but they differ in the model that is used: a 2P normal ogive model (Janssen et al., 2000), a 3PL (Glas & van der Linden, 2003; Sinharay et al., 2003), a generalized partial credit model (Johnson & Sinharay, 2005). Van den Noortgate, De Boeck, and Meulders (2003) use a 1PL model and two approximative estimation methods (PQL from GLIMMIX, and PQL2 from MLwiN).

In summation, from the literature there seem to be three specific reasons for being interested in random item models: the clearly random nature of the items, such as randomly drawn words from a vocabulary; the study of ability change in a longitudinal design with randomly drawn item samples; and modeling item families. More generally, and underlying each of these, is the generalizability potential of random item models. The argument is spelled out in its various aspects by Briggs and Wilson (2007) in their article on the "Generalizability in Item Response Modeling" (GIRM) approach.

Although random item IRT models may be a beginning trend, the vast majority of studies still treats items as fixed. It will be argued here that treating items as random is a general tool, with substantial advantages in various respects to be illustrated further on, reaching beyond the studies reported thus far. Most likely, there are still many other issues where random item IRT models can shed new light.

1.1. The Basic Illustrative Model

For the sake of simplicity, we will deal with the Rasch model for binary responses, in which the logit of the probability of a 1-response is defined as the simple sum of the ability of the person and the easiness of the item, or minus the difficulty:

$$\eta_{pi} = \theta_p - \beta_i, \tag{0}$$

where η_{pi} is $\ln(\Pr(Y_{pi}=1)/\Pr(Y_{pi}=0))$, and Y_{pi} is the response of person p ($p=1,\ldots,P$) to item i ($i=1,\ldots,I$), θ_p is the ability of person p, β_i is the difficulty of item i.

This model can easily be expanded with a degree of discrimination per item, by inserting a parameter α_i as a weight of θ_p, yielding a 2PL model or Birnbaum model, by inserting a guessing parameter, yielding the 3PL, but also by expanding Y_{pi} into a multicategorical variable, either with ordered categories or not, so that models such as the partial credit model and the like can be considered. However, these extensions will not be implemented here, because we first want to explore the rather unknown territory of random items with a rather simple and robust model, such as the Rasch model for binary responses.

2. Why Items May Be Considered Random Some of the Time

Apart from the potential of the random approach to solve problems and to open new perspectives, also two more theoretical reasons may be put forward for why items may be considered random some of the time.

The first reason for treating elements as random is that they are drawn from a population. The *population argument* seems to apply in a natural way for persons. Most populations one can think of, have to be subdivided into much specific subpopulations for the common assumption of a normal distribution to hold. For example, sometimes specifications up to one's ethnicity, gender, education, profession, etc. are necessary. In a well-developed design, the remaining random variation may be rather moderate, which means that the range of the corresponding populations is rather small. In terms of the analogy with items, the person populations are sometimes as specific as item families. An important difference between persons and items is that persons do and items do not pre-exist before they are "drawn." However, there are some preexisting item populations. For example, each language provides us with a vocabulary, and with the advent of computer adaptive testing, items banks are built which may be considered item populations. In the field of educational measurement, the concept of a "universe" or "domain" has been put forward in the context of criterion-referenced measurement (Hively, Patterson, & Page, 1968; Popham, 1978). Furthermore, item generation can be seen as formally equivalent with drawing from a theoretical population. Each generation is a draw. If the items are generated by item writers, they are generated with some concept in mind, perhaps only in an implicit and vague way. On the other hand, when the items are automatically generated, sometimes on-the-fly, an explicit concept is used (Bejar, Lawless, Morley, Wagner, Bennett, & Revuelta, 2003; Embretson, 1999). The concept specifications delineate the populations. One concept can lead to several populations depending on the specifications, sometimes very specific, almost as specific as a single item, for example, as in item cloning (Glas & van der Linden, 2003). The set of possibly generated items is a population.

The second reason for treating elements as random is the uncertainty about the parameters. The *uncertainty argument* leads to the idea of a prior distribution, expressing the possible variation before having more information. After obtaining information from the data, the uncertainty can be reduced, but if a prior distribution is used with unknown parameters (e.g., the variance), then one can estimate the distribution. This distribution is equivalent with the one that expresses the prior uncertainty, and also equivalent with a population distribution as if the elements were random. In a fully Bayesian approach, the parameters of the distribution are themselves provided with a prior distribution, one with hyperparameters.

All this said, random item models can make sense also from a more fundamental, theoretical point of view. The theoretical reasons for random items add to the justification of using a random items approach for more practical reasons, as a useful perspective on various issues in psychometrics to be shown in the following.

3. Three Issues

From a more practical point of view, we will concentrate now on three issues where random item models may help. The first issue is the *measurement of people's ability* using their item responses. Commonly, a fixed item and random person approach is used; the so-called marginal maximum likelihood method:

$$\eta_{pi}|\theta_p = \theta_p - \beta_i \quad \text{and} \quad \theta_p \sim \mathrm{N}\big(\mu_\theta, \sigma_\theta^2\big), \tag{1a}$$

where μ_θ and σ_θ^2 are the mean and variance, respectively, of the ability distribution.

The estimation of this model yields estimates for the item difficulties and estimates for the mean and variance of the person distribution. It has two serious drawbacks given the purpose of its use. First, not the individual persons but the individual items are being measured. An additional step is required to measure the individual persons, assuming the item difficulties are known (e.g., Bock & Mislevy, 1982). Second, the approach lends itself to a generalization over persons for the measurement of the items, whereas in fact a generalization over items is wanted for the measurement of persons. The ideal approach would be one with fixed person effects and random item effects:

$$\eta_{pi}|\beta_i = \theta_p - \beta_i \quad \text{and} \quad \beta_i \sim \mathrm{N}\big(\mu_\beta, \sigma_\beta^2\big), \tag{1b}$$

where μ_β and σ_β^2 are the mean and variance, respectively, of the difficulty distribution.

This *person measurement model* would yield direct estimates of the person abilities and is appropriate to generalize this measurement over items. This does not mean that for other purposes, the model of (1a) cannot be the ideal approach.

The second issue is the *explanation of item difficulties*. A common model for this purpose is the linear logistic test model (LLTM) (Fischer, 1973). The LLTM defines the difficulties as a linear function of item properties specified in the property matrix. These can be item design factors, but in general, any type of item covariates

$$\beta_i = \sum_{q=1}^{Q} \beta_q X_{iq}, \tag{2a}$$

where X_{iq} is the value of item i on property q, and β_q is the weight of property q ($q = 1, \ldots, Q$) to determine the difficulty β_i.

Unfortunately, the LLTM is a rather unrealistic model because it implies that the explanation of the item difficulty is perfect, whereas this is almost never the case. Therefore it helps, just as in a regular regression model, to add an error term, so that the *LLTM with error model* (LLTM + ε) is obtained:

$$\beta_i = \sum_{q=1}^{Q} \beta_q X_{iq} + \varepsilon_i, \tag{2b}$$

where ε_i is an error term with a normal distribution, $\varepsilon_i \sim \mathrm{N}(\mu_\varepsilon, \sigma_\varepsilon^2)$, so that, in fact, the items are treated as random. Strictly speaking, each combination of X-values defines an item population.

The third issue is *differential item functioning* (DIF). DIF means that the probability of giving a correct response to an item, or in general of giving a particular response, is not solely a function of the person's ability, but also of the person belonging to a certain group. The term DIF refers to the differential functioning of an item depending on the group a person belongs to:

$$\beta_{i1} = \beta_{i0} + g_p \delta_i, \tag{3a}$$

where β_{i0} and β_{i1} are the fixed difficulties of item i in the reference group, and in the focal group, respectively, g_p indicates to which group a person p belongs, $g_p = 0$ for the reference group, and $g_p = 1$ for the focal group, δ_i is the fixed DIF value, the difference between β_{i0} and β_{i1}, $\delta_i = 0$ means there is no DIF.

To find out about DIF, one can either make use of indices or of an IRT model. There are two global strategies. The first can be used for both, the second is limited to IRT models. First, following an *anchoring strategy*, one can make assumptions about a subset of the items not showing DIF (anchor items), while investigating the remaining (see an overview of anchoring methods further on). Second, following a *free parameters strategy*, one can either formulate an augmented model without any equality constraints, or one can conduct separate analyses for the focal and reference groups, in both cases, in order to check for which parameters there are differences between the focal group and the reference group. An overview of methods is given among others by Camilli and Shepard (1994), Holland and Wainer (1993), Millsap and Everson (1993), and Teresi (2001).

Both these global strategies have drawbacks. When relying on the anchoring strategy, one depends for the DIF analysis on the quality of the anchor set and corresponding assumptions. When using a free parameters strategy, a lot of parameters are required, and the strategy is more vulnerable to capitalization on chance.

Two way-outs will be described. First, a bivariate approach with random difficulties in both groups can be followed, summarizing the whole of DIF with three parameters: two variances and a covariance. This leads to the *random item profiles model* or RIP model, to be explained later. Second, a binary latent item variable may be introduced to differentiate between DIF items and non-DIF items, leading to an item mixture model. A combination of both, called the *random item mixture model* or RIM model, would lead to the following formulation:

$$\begin{aligned} &\beta_{ig} = (1 - \alpha_i g_p)\beta_{i0} + \alpha_i g_p \beta_{i1} \quad \text{and} \quad (\beta_{i0}, \beta_{i1}) \sim \text{BVN}(\mu_{\beta_0}, \mu_{\beta 1}, \Sigma_{\beta_0\beta_1}), \\ &\alpha_i \sim \text{Bernoulli}(\pi_\alpha), \end{aligned} \tag{3b}$$

where β_{ig} is the difficulty of item i in group g, α_i is a latent binary variable indicating whether an item shows DIF ($\alpha_i = 1$) or not ($\alpha_i = 0$), μ_{β_0} and μ_{β_1} are the mean difficulties in the reference group and the focal group, respectively, $\Sigma_{\beta_0\beta_1}$ is the covariance matrix of the difficulties. Note that the bivariate distribution is conditional on $\alpha_i = 1$.

The model in (3b) makes use of two kinds of random item variables: continuous ones (β_{i0} and β_{i1}), and a binary one (α_i).

There might be other issues that can be approached with a random item IRT, but we will focus on these three. They are chosen to be representative for some of the most important subjects in the field of psychometrics: measurement per se, explanatory measurement, and troubleshooting measurement. *Measurement per se* is an important subject, especially because decisions are being based on measurement results, such as in education. Explanation is another important subject because explanation is the main objective of science, and the basis of all practice. If what is established cannot be explained, one can at best come up with a trial and error approach for a remedy. *Explanatory measurement* is the combination of measurement and explanation, so that what is measured can be explained at the same time (De Boeck & Wilson, 2004). *Trouble shooting measurement*, or the identification of problems through measurement, and most importantly problems with the quality of measurement, is another important subject. Most trouble shooting approaches are heuristic, using indices with known or unknown statistical properties. The use of indices for DIF is a common practice. Modeling is often not a good alternative for troubleshooting because most models are not directed to the identification of problems; they rather deal with the regular case. But when the measurement model is shaped after the problems to be identified, also a measurement model may help for trouble shooting.

4. Verbal Aggression Data Set

One data set will be used for all three kinds of subjects, based on De Boeck and Wilson (2004). It is a 316 persons by 24 items binary data set, obtained after a dichotomization of ordered-category responses (0 = no, 1 = perhaps, 2—yes, with 1 and 2 recoded as 1). All items consist of a stem describing a frustrating situation, and a verbal aggression response part describing how people could respond to the situation in question. The items are written to fit a $2 \times 2 \times 3$ design with two replications within each cell.

For the item stem, a two-valued factor is used: others to blame vs. self to blame. For example, "a bus fails to stop for a passenger" is a frustrating situation for the passenger, but it is not the passenger's fault; instead it is the bus driver's fault. On the other hand, when "someone is about to enter a grocery store just when it closes," the person is most likely to be frustrated, but it is not the fault of the storekeeper when the closing hour is respected; instead the client should have known better. In total, four situations were used:

1a. A bus fails to stop for me.
1b. I miss a train because the clerk gave me faulty information.
2a. The grocery store closes just as I am about to enter.
2b. The operator disconnects me when I used up my last 10 cents for a call.

The first two situations are other-to-blame situations, and the latter two are self-to-blame situations. The two situations of the same type (a and b) are considered as replications.

For the verbal aggression responses, a person may give to the situation in question, two factors are varied: three kinds of verbal aggressive behaviors ("cursing," "shouting," "scolding"), and two behavior modes ("wanting," "doing"). The combination of these two factors yield response formulations such as "I would curse" (for doing), "I would want to curse." (for wanting)

The full item contains an item stem (one of the four frustrating situations), and a response to the situation (one of the three behaviors combined with one of two behavioral modes), so that given there are two items of each kind, in total 24 items are obtained, while the item design is of a $2 \times 2 \times 3$ type.

In total, 316 persons have responded to the items: 243 women and 73 men, all of them first-year psychology students of a Dutch speaking Belgian university. The 24 items were presented in Dutch, the native language of all the participants.

5. Four Rasch Models

Depending on whether the persons and the items are either treated as random or fixed, four different kinds of Rasch models can be defined: (1) the fixed persons—fixed items Rasch model (FPFI Rasch), (2) the random persons—fixed items Rasch model (RPFI Rasch), (3) the fixed persons—random items Rasch model (FPRI Rasch), and (4) the random person—random item Rasch model (RPRI Rasch) (see Table 1).

TABLE 1.
Four Rasch models depending on the random or fixed nature of persons and items.

	Items	
Persons	Fixed	Random
Fixed	FPFI Rasch	FPRI Rasch
Random	RPFI Rasch	RPRI Rasch

Two of these models are regularly used: the FPFI Rasch model and the RPFI Rasch model. The former is the joint maximum likelihood (JML) version of the Rasch model, and the latter is the marginal maximum likelihood (MML) version of the Rasch model. It is well known from the literature that the JML estimation method leads into consistency problems, and that a correction of the person parameters is needed to obtain better estimates (Andersen, 1980). This is why the MML approach is preferred.

Usually these two approaches are described as estimation methods, but they rely on different model formulations. The reason why they are considered estimation methods in the first place is because when the focus is on the estimation of item parameters, then assumptions regarding the persons may be considered instrumental to arrive at item parameter estimation. The third classical estimation method is conditional maximum likelihood (CML), which is a method without making any assumptions regarding the persons, based on the given that for the Rasch model, the sum score is a sufficient statistic for the person parameters. Again, the same priority is given to the estimation of the item parameters. Apart from the items being considered in a more natural way as fixed entities, the fact that the data contain more information about the item parameters than about the person parameters is perhaps the main reason why one has concentrated on items for the estimation.

Two of these models are not regularly used: the FPRI Rasch model and the RPRI Rasch model. The first of these is the earlier mentioned person measurement model. Its estimation can follow the well-known MML method, but with changing roles between persons and items. The second is a crossed random effects Rasch model, with both, random person effects and random item effects. It requires a different approach to estimation as will be explained.

Instead of reducing the issue of choosing between the four to an issue of estimation, the choice should depend in the first place on one's purpose. Generalization and explanation are two important purposes that would lead to the use of random parameters. A random persons approach is to be recommended if one wants to generalize the item measurements over persons to the population of persons under consideration, such as when building an item bank for computer adaptive testing. A random persons approach is to be recommended also if one wants to explain the person variation though external person covariates because the random approach allows for an error term. In a similar way, if one wants to generalize person measurements over items, or if one wants to explain item difficulties, a random item approach is to be recommended. Measurement of individual elements is an important purpose that would lead to the use of fixed parameters. If interested in the individual items or individual persons, the items and persons, respectively, should be treated as such, and included in the model with fixed effects.

The choice of a model, one out of the four, has consequences for the goodness of fit of the model, for a possible shrinkage of the parameter values, and for the standard error values. Furthermore, the estimate of the intraclass correlation depends on the model. All this will be illustrated for the verbal aggression data set, although none of the models may seem perfectly appropriate, for one or more of the following reasons. The models do not take into account that men and women may be two different populations, and in a similar way, the item design is not taken into account. There might be also other reasons, such as the difficulties differing between men and women.

5.1. Model Estimation

Three of the four Rasch models are estimated with the lmer function of lme4 in R (Bates, Maechler, & Dai, 2008), and all four are estimated with WinBUGS, a Markov Chain Monte Carlo (MCMC) approach (Lunn, Thomas, Best, & Spiegelhalter, 2000). The one model that could not be estimated with lmer is the FPFI model, because lmer requires at least one random effect, and for that model the glm function from R was used. For the estimation of the standard errors, lmer makes use of the posterior distribution of the parameters in question (command mcmcamp)

which can be derived as follows. Because the likelihood is known, assuming a certain prior, the posterior can be derived. For the fixed effects parameters, a locally uniform prior is chosen. For the variance-covariance matrices of the random effects, also a locally noninformative prior is used, an inverse Wishart distribution in the neighborhood of the estimates.

The second program, WinBUGS version 1.4.3 was ran with five chains with a length of 5,000 each. The priors for the random effects were $N(0, \sigma^2)$, with σ^2 distributed as an inverse gamma (1, 1), and the priors for the fixed effects were $N(0, 1/0.05)$, but the results were identical when 0.1 was used as the precision value (and hence, 1/0.01 as the variance). These options for WinBUGS will be used for all applications, also in the following sections. In order to check convergence, the $\hat{R}$ index proposed by Gelman and Rubin (1992) is used. The values are not reported because they never exceeded the critical value of 1.1.

For the estimation of the RPRI Rasch model, two more methods were used: the xtmelogit command from Stata (StataCorp, 2007), and a newly developed alternating imputation posterior (AIP) algorithm with adaptive Gauss–Hermite quadrature (Cho & Rabe-Hesketh, 2008). The RPRI model is a model with crossed random effects and, therefore, it requires integration for each item and for each person. Two of the four estimation methods are based on an approximation of the integrand and two are based on an approximation of the integral. See Tuerlinckx, Rijmen, Verbeke, and De Boeck (2006) for a discussion of the various approximation methods. Both, xtmelogit and lmer make use of the Laplace method (Thierny & Kadane, 1986), which is a quadratic approximation of the log of the integrand at its mode. The Laplace method is expected to work well if the cluster size (number of items in the RPFI model, and number of persons in the FPRI model) is rather large. It is known, however, that all estimates are somewhat less extreme when an approximation to the integrand is used. The other two approaches, WinBUGS and AIP, are based on an approximation of the integral, one is a sampling method (MCMC with WinBUGS), and the other is an adaptive Gauss–Hermite quadrature approach (AIP).

The main idea of the AIP algorithm is to divide the total model into submodels (called wings) for persons and for items, and to alternate between the two WINGS until convergence. The algorithm iterates between an item wing in which the item mean and variance are estimated for given person effects and a person wing in which the person mean and variance are estimated for given item effects. The estimation is based on maximum likelihood and adaptive quadrature. The person effects used for the item wing are sampled from the posterior distribution estimated in the person wing and vice versa.

5.2. *Results*

5.2.1. Goodness of Fit We will not investigate the absolute goodness of fit because the models are used for illustrative purposes, but instead the relative indices AIC and BIC (although strictly speaking, these indices don't apply for methods which are based on an approximation of the integrand, such as lmer). Table 2 shows the deviance and the AIC and BIC results for all four models (the function glm from R for the first, and the function lmer for the other).

It is by definition the case that the FPFI has the lowest deviance because it is the most flexible model, with a parameter for each of the items and for each of the persons. Because the number

TABLE 2.
Goodness-of-fit indices of the four models estimated with lmer.

Rasch models	# df	Deviance	AIC	BIC
FPFI	48	7070	7166	7499
FPRI	26	7196	7248	7428
RPFI	25	8078	8128	8302
RPRI	3	8203	8209	8230

of persons is larger than the number of items, it is also a quite normal result that the FPRI model has the second lowest deviance, followed by the RPFI model, and the RPRI model, which has the highest deviance.

Like in all Rasch models, one degree of freedom is spent on an identification constraint. The FPFI model needs in theory $P + I - 1$ parameters, but following the Rasch model, only $I + 1$ different values for the fixed person parameters can be obtained (because there are only $I + 1$ sum scores), so that in practice only $(I + 1) + I - 1 = 2I$ parameters are used. The FPRI model needs in theory $P + 1$ parameters and in practice only $(I + 1) + 1 = I + 2$ parameters. The RPFI models needs $1 + I$ parameters. Finally, the RPRI model needs only three parameters: one mean and two variances, the mean of the persons, or of the items being set equal to zero for reasons of identification.

Because of its large flexibility and the still relative low numbers of parameters, the FPFI model has the lowest AIC value. The FPRI model is the second best in terms of AIC. Given that the numbers of parameters are still much larger for the FPFI model than for the FPRI model, and that the deviance is not much larger, it is not surprising that the FPRI model has the best relative goodness of fit in terms of the BIC. The RPFI model has the highest BIC because its deviance is not much lower than that of the RPRI model, while the latter uses much more parameters. Note that the FPRI model was considered here to be the ideal model for measurement, while the RPFI model is the most popular one. However, in terms of goodness of fit (deviance, AIC, and BIC), the FPRI model does clearly better than the RPFI model.

Two factors play in the goodness of fit of the four models. First of all, one may expect the goodness of fit of a model with fixed persons/items to be better the more the distribution of the persons/items deviates from the normal distribution. This factor depends on the data. Second, for the information indices, AIC and BIC, the number of parameters plays an extra role. In this regard, models with fixed effects have a handicap (a larger penalty) in comparison with the models with random effects. This factor depends on the size of the data set, independent of the data. That the fixed vs. random effects of the persons have a larger effect than the fixed versus random effects of the items, may perhaps not be generalized to other applications.

5.2.2. Shrinkage Effects Goodness of fit has an effect on the scaling of the effects because in IRT the effects are expressed relative to the unexplained variance, or more precisely, relative to the standard deviation of the error term (Snijders & Bosker, 1999). In a logistic model, the unexplained variance is the logistic error variance. The effects are expressed relative to the standard logistic variance, which is 3.290. However, when the logistic distribution is approached with a normal distribution, the corresponding variance 2.892 (the square root of which is the well-known scaling factor D or 1.70), but see Savalei (2006) for a Kullback–Leibler alternative, with $D = 1.75$. If a model is incomplete because some effects are not included in the model, these nonincluded effects increase the error variance, which is standardized, so that in fact the scaling of the included effects shrinks, unless a dispersion parameter is included.

In order to check for this effect, the ability estimates and difficulty estimates of all four models are pairwise regressed one onto the other, both for lmer and WinBUGS. It was found that the difference in goodness of fit (deviance), was highly predictive of the slope. When effects from a better fitting model were regressed on those of a poorer fitting model, the slope was smaller than one (except in some of the cases, when the difference was small). For example, regressing the glm abilities from the FPFI model on the lmer abilities of the other three models, the slopes were 1.010 (FPRI), 0.823 (RPFI), and 0.809 (RPRI), nicely in line with the differences in the deviance (126, 1008, 1133). This result was obtained also for WinBUGS and for the item difficulties, but the effects are larger for WinBUGS and for abilities in comparison with difficulties.

The shrinkage affects also the variance estimates. The estimates are larger if the deviance is smaller. For example, the lmer estimates of the difficulty variance are 1.419 for the FPRI model

and 1.280 for the RPRI model. The effect is smaller for the ability variance: 1.902 for the RPFI model and 1.887 for the RPRI model when lmer is used. This is again in line with the difference in deviance. The difference between the FPRI and RPRI models is 1,007, while the difference between the RPFI and RPRI models is only 125. The same kinds of effect show with respect to the standard errors of the estimates, and again the shrinkage is a function of the goodness of fit.

The choice of a model is not without consequences for the size of the effects, due to the fact that all effects are scaled relative to a standardized error. One should therefore not be surprised to see rather large differences in estimates when going from one model to the other. These differences must be attributed to differences in goodness of fit, rather than to the estimation method as one may be inclined to when an estimation perspective predominates, as for JML and MML.

5.2.3. Variance Components An interesting consequence of treating elements as random is that the variance components and an intraclass correlation can be determined. The only variance component for the FPFI model is the error variance (the standard logistic variance), so that the intraclass correlation cannot be derived. For the FPRI model, the item variance component is estimated, and for the RPFI model, the person variance component is estimated. Finally, for the RPRI, three estimated variance components are available: the error variance component, the person variance component, and the item variance component. Various kinds of intraclass correlations can be calculated: for the persons or for the items, the other mode being fixed or random, and for two elements, ICC(1) (the reliability of one item), or for the sum of all elements, ICC(k) (the reliability of the sum) (McGraw & Wong, 1996; Shrout & Fleiss, 1979).

We concentrate here on the ICC(k) or the reliability of the sum. The intraclass correlations reported in the test literature always concern the persons, and refer to the reliability of the measurement of persons. This makes sense, of course, because one is interested in the measurement of persons in the first place. The items are "only a tool." In general, the intraclass coefficient for items is less relevant, but perhaps not irrelevant in the context of building an item bank.

The most complete model with respect to variance components is the RPRI model. The variance estimates are given in Table 3 for lmer, xtmelogit, WinBUGS, and AIP. They are remarkably similar throughout all four programs, and only for the person variance slightly smaller when an approximate method is used. The lmer results will be used for the ICC, but, the ICCs are nearly identical for all four methods. The three variance components are as follows: 1.887 (persons), 1.280 (items), and 2.892 (error). For the error component, the normal approximate is used, because the other two distributions (persons, items) are normal. This means that 31.1% of the variance is due to persons, and 21.1% is due to items, whereas almost half is error variance. The total variance (100%) refers to a theoretical underlying continuous variable V_{pi} which is the sum of both random effects and an error term. In order to obtain binary responses following the Rasch model, this V_{pi} must go through a dichotomization process with as a cut-off $V_{pi} = 0$ ($Y_{pi} = 1$ iff $V_{pi} \geq 0$). The corresponding ICC(k) for V_{pi} is 0.916. When the items are treated as fixed, following the RPFI model, the corresponding ICC(k) is 0.940. These results are also confirmed when the corresponding normal-ogive models are used. However, when the coefficient alpha is calculated based on the binary data, the result is 0.876. The difference with the ICC(k) values

TABLE 3.
Variance estimates of the RPRI Rasch model using four methods.

Method	Person variance	Item variance
lmer	1.887 (0.19)	1.280 (0.47)
xtmelogit	1.886 (0.19)	1.276 (0.38)
WinBUGS	1.911 (0.19)	1.289 (0.40)
AIP	1.901 (0.19)	1.277 (0.38)

just reported is that the latter refer to the underlying continuous variable V_{pi}, whereas the 0.876 refers to its dichotomization, which is less informative. Another explanation for the difference is that coefficient alpha is lower when there is an interaction between persons and items. However, if that is the case, the Rasch model is contradicted.

The corresponding ICC(k) for the items is 0.914, and if the persons are fixed, the value is 0.922. These values may be considered high, but they rely on the assumption that the distribution of difficulties is normal, which is not a desirable feature for all kinds of item banks. If one wants to cover the ability in an equally good way through its whole range, then a rather uniform distribution is desirable in the range between plus and minus three standard deviations.

Models with random effects have the asset that they allow for the estimate of variance components and ICCs, also for the items if the items are treated as random. The latter may be especially relevant for applications in which one relies on an item bank. One should keep in mind, though, that by definition the ICC is larger when the other mode is treated as fixed, although the difference may be small, as shown in the present application.

5.3. Discussion and Conclusion

It was argued earlier that for the measurement of persons per se, the *person measurement model* or FPRI model is the better one, because it concentrates on individual persons and has the potential to generalize over sets of items. The model did quite well also in the application with a clearly better goodness of fit than the common RPFI model, and in terms of the BIC, it was the best of all four models. Strictly speaking, this result may not be generalized because it depends on how far the person distribution differs from the normal distribution, and on the number of persons and items, but given that there are always far more persons than items, it may be expected in most cases indeed. Interestingly, the FPRI model offers also the possibility of assessing the reliability of the item difficulties in a rather easy way, using the intraclass correlation. All these assets contribute to the value of the FPRI model.

However, the purpose is not always measurement per se of persons. Just as for items, there are often good reasons to treat the persons as random, and items as fixed. As has been shown, choosing for one of the four models has implications, for the goodness of fit, for the scaling of the effects and their standard error, and for the value of the ICC.

The doubly random model, RPRI, is considered a challenge for estimation. In the present application, four clearly different estimation methods yield very similar results. In theory, the AIP and the MCMC methods should be the better ones, but in this application their advantage is certainly not outspoken. Apart from the methods used here, also GLIMMIX and MLwiN can be used for the estimation of doubly random models (Van den Noortgate et al., 2003).

6. The Linear Logistic Test Model with Error

The reason for being interested in a LLTM with an error component is twofold. First, it is rarely the case that the item parameters can be perfectly explained from the item properties, because both the substantive theory behind the model is not perfect, and because the difficulty may actually be a random variable. This is especially the case if more than just one item is used for each cell in the design, as is the case for the data set under consideration with 24 items for 12 cells in the design. An error term as in (2b) can then account for the discrepancy between the freely estimated difficulty and the LLTM-estimated difficulty.

Second, it is possible that the error term is larger in one part of the design than in another. For example, going from “wanting” to “doing,” a lot of additional factors may start to play a role, beyond and independent of the item properties that are hypothesized to play (the design

factors). The hypothesis of additional factors not covered in the design can be tested by allowing for a heteroscedastic model, with a different error term depending on whether doing or wanting is concerned. The data set is perhaps not very appropriate because of its limited size, with only 12 doing items and 12 wanting items, but the application is meant for illustrative reasons only.

An error component as in (2b) makes the item difficulty prediction into a regular regression model, with an error term, just as in a latent regression approach used for the person parameters, using design factors for the persons (Adams, Wilson, & Wu, 1997; Verhelst & Eggen, 1989; Zwinderman, 1991).

6.1. Model Estimation

The coding scheme for the item properties is the same as used by Janssen, Schepers, and Peres (2004). Dummy coding is used for "doing" (= 1) versus "wanting" (= 0), and "other-to-blame" (= 1) versus "self-to-blame" (= 0). For the three behaviors, contrast coding is used with two factors: (1) a blaming factor, with "curse" and "scold" (= 1/2) as blaming behaviors, contrasted with "shout" (= −1), which is in Dutch ("het uitschreeuwen") not really a way of blaming someone else; (2) an expression factor, with "curse" and "shout" (= 1/2) as expressive behaviors, contrasted with "scold" (= −1).

The data have been analyzed before with both the models formulated in (2a) and (2b) by Janssen et al. (2004). For the LLTM with error (homoscedastic), a MCMC approach implemented in the software of the authors was used, whereas both, the same MCMC approach and NLMIXED (SAS Institute, 1999), were used for the regular LLTM. In the present application, three models (LLTM, homoscedastic LLTM + ε, heteroscedastic LLTM + ε) are estimated with lmer and WinBUGS, using the same options as in the earlier applications.

6.2. Results

6.2.1. Goodness of Fit The deviances of the three models as estimated with lmer are as follows: 8238 (df = 6), 8150 (df = 7), and 8147 (df = 8) for the LLTM, homoscedastic LLTM + ε, and heteroscedastic LLTM + ε, respectively. The corresponding AIC and BIC values are: 8250 and 8292 (LLTM), 8164 and 8213 (homoscedastic LLTM + ε), and 8163 and 8218 (heteroscedastic LLTM + ε). The LLTM models with error do clearly better than the regular LLTM, but it is hard to differentiate between the two models with error. As expected, the error term seems larger for the do items than for the want items, the respective estimates being 0.21 (0.17) for doing and 0.04 (0.09) for wanting, but the large standard errors (within parentheses) do not allow for an interpretation of the difference. The standard error is given within parentheses also in the following. The number of items is too low for a conclusion to be drawn about homoscedasticity versus heteroscedasticity. In the following, only the homoscedastic variant will be reported. The item variance estimate for the homoscedastic model is 0.16 (0.06), and the standard error of 0.06 is smaller than for the heteroscedastic model.

6.2.2. Comparison of Estimates The estimates from lmer and WinBUGS for the regular LLTM and the LLTM with homoscedastic error are shown in Table 4. First of all, these results perfectly agree with those obtained by Janssen et al. (2004). The interpretation must be that people are less aggressive in what they would do than in what they would want to do, which is a clear and expected inhibition effect (doing makes the items more "difficult"). Verbal aggression is more likely when someone else is to blame (someone else to be blamed makes the items "easier"), and both, blaming and expressing one's frustration are more likely than not blaming the other person and not expressing one's frustration (blaming items and expressing items are "easier").

TABLE 4.
Estimates of the item property effects for the regular LLTM and the LLTM + ε.

	LLTM lmer WinBUGS	LLTM + ε lmer WinBUGS
Do	0.68 (0.06) 0.67 (0.06)	0.71 0.15) 0.69 (0.17)
Other-to-blame	−1.03 (0.06) −1.03 (0.06)	−1.05 (0.15) −1.03 (0.19)
Blaming	−1.35 (0.05) −1.36 (0.05)	−1.40 (0.12) −1.41 (0.16)
Expressing	−0.70 (0.05) −0.70 (0.05)	−0.70 (0.12) −0.71 (0.16)
Intercept	0.32 (0.05) 0.31 (0.05)	0.33 (0.15) 0.33 (0.17)

Second, the results are highly similar for all estimation methods used: the MCMC method as programmed by Janssen et al. (2004), NLMIXED for the regular LLTM, lmer, and WinBUGS. This adds to the robustness of the results.

Third, in general, the estimates of the effects are slightly larger for the LLTM with error (LLTM + ε) than for the regular LLTM. The difference can again be explained by a scaling effect because the LLTM + ε has a smaller deviance. The effect is not large, the correlation between the Rasch difficulties and the LLTM reconstructed difficulties is 0.94, and hence the improvement using the LLTM + ε is only minor, although worth of considering. Another way of looking at the predictive value of the item properties is to compare the error variance estimate of 0.16 obtained from the LLTM + ε, with the item variance estimate from the RPRI model, which is 1.28 using lmer. Hence, the item properties reduce the item variance with 87.5%, or in other words, 87.5% of the original variance is explained. This corresponds exactly with a correlation coefficient of 0.94, as derived earlier in another way.

Fourth, the standard errors of the estimates are clearly larger for the LLTM + ε in comparison with the regular LLTM, and both methods, lmer and WinBUGS agree on this finding. The explanation cannot be a scaling effect because the scaling effect is only minor, and neither can it be the extra uncertainty when using a Bayesian method (which leads to somewhat larger standard errors). The larger standard errors must thus stem from using an error term in the LLTM.

6.3. Discussion and Conclusion

The LLTM + ε corresponds better with a (latent) regression approach, and has therefore a conceptual advantage. Apart from its conceptual advantage, two other assets should be mentioned, based on the application: (1) The LLTM with error yields a better goodness of fit, although the gain is rather minor in this application, but it is not negligible either. The usefulness of a heteroscedastic error term could not be illustrated with this data set, probably because the item set is too small. (2) The various estimation methods yield about the same results, which suggest that it is quite feasible indeed to estimate a LLTM with error. This is not surprising since a similar approach for persons, called a latent regression approach (e.g., Zwinderman, 1991), seems to work well. In the present application, the item covariates have a very strong predictive value, and it would be of interest to investigate the model and its estimation also for the case with weaker item covariates because especially in those cases the error term would be badly needed.

The LLTM + ε has a price to be paid, in the form of much larger standard errors. But when the regular LLTM is not the true model because there is some random variation left, the effects as estimated in the regular LLTM may look more robust than they actually are. Effects that seem significant using a regular LLTM could as well be not significant when an error term is used. It is an important issue which statistical conclusion should be trusted. Given the strength of the covariate effects in the present application, it is not an issue here, but it might be in other applications. The LLTM is almost always a misspecified model. Eliminating the error term as in the regular LLTM means one actually eliminates the uncertainty associated with the unexplained item variance.

7. Differential Item Functioning

The issue of differential item functioning (DIF) can be formulated as follows. Which of the items, if any, has an IRF that differs between (1) a group of interest, such as a minority group or any other group one would be interested in, and is therefore called the focal group, (2) and a reference group, taking into account that the mean ability may differ depending on the group? Given the concentration on Rasch models, we will limit the study to difficulty as the sole aspect in which IRFs can differ. This is of course a limitation, and it restricts the present study to uniform DIF, but it is an initial step to explore several new approaches, to be investigated further for other aspects of the IRF in the future.

7.1. A Bivariate Difficulty Distribution Approach

DIF is studied either in a nonparametric way, without any modeling, or it is studied in a parametric way, based on modeling. However, in both cases, DIF is seen as a discrete event. An item is flagged as DIF or it is not. A discrete view is perhaps not a realistic view on DIF, for the following reasons:

1. DIF values differ depending on the items. In many cases, the DIF index does not show an elbow when the values are ordered from high to low. The decrease from the DIF items to the non-DIF items is often not abrupt. Smaller values of these indices do not mean there is no difference. It seems very unlikely that the degree of DIF drops to zero or almost zero when going from DIF items to non-DIF items. It seems plausible that in many cases DIF decreases gradually. This graduality may be better in line with a random effects approach than with a fixed effect approach and a clear cut between DIF and non-DIF.
2. The items may differ in the degree to which they are characterized by a feature that makes the difference between the two groups. For example, math problems often require some verbal comprehension, and the verbal comprehension ability may differ between the focal group and the reference group. It is hardly possible to keep the degree of required verbal comprehension perfectly constant over the items, and it is also rather unlikely that math problems fall nicely into two categories for their loading with verbal comprehension (one with, one without such a loading). A gradual approach with random differences between the items may be more appropriate to capture such a reality.
3. Small DIF values do not make a chance with the classical DIF approach because the method in question often does not have the power to detect these small values, so that only the larger ones are identified. Statistical significance or rules of thumb are used as a decision criterion, so that DIF may look like a discrete phenomenon restricted to a few items, whereas in fact its nature may be gradual and global instead. Using a discrete decision criterion does not make reality discrete.

A more gradual view on DIF and its possible random nature leads into a model with random item profiles in the focal and the reference group (RIP model). This bivariate approach is nothing more than a RPRI Rasch model for both groups. But because we now have two random difficulties, one for each group, a bivariate distribution of item difficulties is involved (see (3b)):

$$(\beta_{i0}, \beta_{i1}) \sim \text{BVN}(\mu_{\beta_0}, \mu_{\beta 1}, \Sigma_{\beta_0\beta_1}). \tag{4}$$

In this way, DIF is no longer associated with a particular subset of items, but is a general and gradual item phenomenon instead. DIF can be tested comparing the univariate and the bivariate models. The degree of overall DIF can be expressed in the correlation and the difference between two variances. Important advantages of the approach just described are that it allows for overall DIF without a serious increase in the number of parameters, and that it can easily be extended into a multivariate approach if there is more than one focal group, so that a multivariate instead of a bivariate RIP model is obtained.

7.2. Simulation Study

Using a RIP approach does not prevent us from still trying to identify DIF items in case DIF is in fact a discrete event. This will be shown in a simulation study where the RIP approach is used together with some more traditional approaches. The RIP approach for flagging DIF items consists of fitting the bivariate RIP model and to derive ML estimates for the difficulties in both groups, followed by a robust regression and the identification of outliers, in order to identify DIF items.

The reason for thinking of robust methods is twofold. First, these methods can solve the linking issue. The estimation of the group difference is not affected by outliers when using a robust method. DIF items are outliers in the scatter plot of difficulties in the focal group versus difficulties in the reference group, on the condition that the DIF items are a minority. This leads to a certain linking that differs from the more common ANOVA linking, with mean difficulties of zero in both groups. Second, the ANOVA linking yields DIF inflation when DIF is asymmetrical, whereas the robust approach does not. DIF is asymmetric if it is restricted to a subset of the items showing a mean difference in difficulty between the two groups. Take as an example five items with the following difficulties -2, -1, 0, $+1$, and $+2$ in the reference group, and -1, 0, 0, $+1$, and $+2$ in the focal group, and with no difference in ability between the two groups. As can be seen, the first two items are DIF items, and their mean differs between the two groups. However, when an ANOVA based linking is used, with mean difficulties of zero, the focal group will be seen as less able, with a mean that is 0.40 lower than that of the reference group, and the difficulties in the focal group would seem to be -1.40, -0.40, -0.40, 0.60, 1.60, and, therefore, all items would look like DIF items. This is an artificial DIF inflation due to the kind of linking (Wang, 2004). As can be seen in the example, the size of the true DIF is reduced, and DIF is created where there is not any. This may lead to both, more false negatives (DIF item not flagged as such) and more false positives (non-DIF items flagged as DIF).

The robust regression method is much less vulnerable to this inflation. It gives no, or only a very small weight to the first and second items, so that the ability difference between the two groups is not much influenced by the DIF items. In the simulation study, five different methods to detect DIF will be used.

The robust procedure consists of two steps after the estimation of the bivariate RIP model and the corresponding difficulties (all with lmer in the simulation study): In *step one*, the focal group difficulties are regressed on the reference group difficulties. The regression is based on a minimization of Tukey's biweight function ρ and is therefore not a regular kind of least squares minimization. In fact, an iterative reweighted least square method is used for the minimization,

such that the degree of extremeness of an observation has no effect. The function ρ-based estimator is an M-estimator (Rousseeuw & Leroy, 1987) and is implemented in the function rlm from the MASS library in R. The reason for using the robust regression is to prevent that the regression line is affected by asymmetric DIF (e.g., all DIF items being more difficult or easier in the focal group). DIF items are in fact outliers, and if the DIF is asymmetric, the intercept in a regular regression would be affected independent of the true impact (difference in ability level between the two groups). Two kinds of regression were used, one with a slope of one, and another with the slope as a parameter. On average, the slope-one regression did slightly better for DIF identification, so that only the results of that method will be reported.

In *step two*, a robust confidence interval is determined for the distance to the regression line. The method used is the MCD approach (robust estimation of the covariance matrix by the "minimum covariance determinant") (Rousseeuw & van Driessen, 1999). The reason for using robust confidence intervals is to prevent that the confidence intervals are stretched by the DIF items (outliers in the scatter plot). Items with a distance from the regression line that exceed the 95% confidence interval are flagged as DIF items. If the regression slope is one, step one of the procedure (the regression step) may be skipped. The method based on estimating the RIP model followed by a robust analysis is called here the *RIP* + *rob* method.

A simpler variant of this approach would be to go through the two robust steps, but with the logit of the proportions of success in both groups, and hence without any modeling. This method is very similar to the delta plot method (Angoff & Ford, 1973) and its regression-based adapted version (Chen & Henning, 1985), apart from two aspects. First, the probit transformation is used for the delta plot method instead of the logit transformation, and second, the delta plot method does not make use of robust methods. The delta method is found to be a suboptimal method in the literature (Ironson, Homan, Willis, & Singer, 1984; Shepard, Camilli, & Williams, 1985), but a robust variant has never been investigated. The robust approach applied to the logit of the proportions of success is called here the *logit*(p) + *rob* method. In contrast with the RIP + rob method, it is a nonparametric method.

Three other methods will be used in the simulation study: two nonparametric methods based on the sum of scores, and one parametric method. They share all three that each item is investigated separately assuming the other items are non-DIF items (all-other anchoring). This assumption is a rather common one, but the risks of this assumption are somewhat compensated by an iterative procedure, omitting items from the sum score one by one. First, the item with the strongest indication of DIF is omitted, and the method is reapplied, and so on (Millsap & Everson, 1993). Here, the methods will be used in a non-iterative way.

The first traditional method is the *Mantel–Haenszel method*, abbreviated as *MH* (Mantel & Haenszel, 1959). The MH χ^2 statistic to determine whether the non-DIF null hypothesis should be rejected is based on a contingency table of item $\times$ group $\times$ sum score. Like for the previous two methods, an α-level of 0.05 will be used.

The second traditional method is the *standardization method*, based on the *STD P-DIF* statistic. The statistic is the weighted sum over the score groups of the differences in proportion of success between the focal and the reference groups, weighted with the proportions represented by the sum score groups of the focal group (Dorans & Kulick, 1986). Dorans and Holland (1993) found that the results of the standardization method are in close agreement with the MH results. Much depends on the critical value that is used. Therefore, different thresholds for DIF identification are tried out. Among the critical values of 0.05, 0.06, 0.07, 0.08, 0.09, 0.10, the threshold of 0.08 gave the best results on average and most similar to the MH method. Therefore, the 0.08 will be used to report the results.

The third traditional method is the *likelihood ratio method* (Thissen, Steinberg, & Gerrard, 1986), abbreviated here as *LR*. The method will be applied here item by item. The LR test is applied using two models: one with equal item difficulties and another with group dependent

TABLE 5.
Error rate in DIF identification for five methods, depending on the distribution of the items, the overall difference in ability, and symmetrical versus asymmetrical DIF.

	DIF identification method	Difficulties N(0, 1)	Difficulties U[(−2, 2])
Symmetrical DIF	MH	3.00%	3.50%
equal mean abilities	STD P-DIF[a]	3.75%	4.00%
	LR	3.75%	5.25%
	Logit(p)+rob	4.00%	4.75%
	RIP+rob	3.25%	4.75%
Asymmetrical DIF	MH	13.00%	12.25%
equal mean abilities	STD P-DIF[a]	12.75%	13.00%
	LR	15.00%	16.00%
	Logit(p)+rob	3.75%	3.25%
	RIP+rob	5.00%	2.75%
Symmetrical DIF	MH	4.00%	5.75%
unequal mean abilities	STD P-DIF[a]	7.50%	9.50%
	LR	4.25%	5.50%
	Logit(p)+rob	3.75%	6.25%
	RIP+rob	3.00%	5.75%
Asymmetrical DIF	MH	13.75%	15.25%
unequal mean abilities	STD P-DIF[a]	16.00%	15.75%
	LR	16.25%	17.50%
	Logit(p)+rob	5.50%	6.75%
	RIP+rob	4.75%	6.25%

[a]The critical values are 0.08 and −0.08.

difficulty parameters for item i. The decision criterion is a LR test outcome that is statistically significant based on an α-level of 0.05.

Data were simulated for 20 items and 500 persons, following the Rasch model. Three factors were varied:

(1) the distribution of the item difficulties in the reference group (β_{i0}): either normal, N(0, 1); or uniform, U([−2, +2)];
(2) four DIF items, either with symmetrical DIF values $(-0.8, 0.8, -1.2, 1.2)$, or with asymmetrical DIF values $(0.8, 0.8, 1.2, 1.2)$;
(3) mean abilities of either zero in both groups ($\mu_{\theta_0} = \mu_{\theta_1} = 0$), or zero in the reference group, and one in the focal group ($\mu_{\theta_0} = 0, \mu_{\theta_1} = 1$).

For each of the 8 combinations, 20 data sets were generated.

The results are reported in Table 5. Three conclusions may be drawn from these results. First, the three traditional methods, but not the two robust methods, are vulnerable to distortion when DIF is asymmetrical. Both, the percentage of false positives (DIF inflation) and the percentage false negatives increase drastically when DIF is asymmetrical, except for the two robust methods. The increase of both types of errors is expected, as explained earlier. Second, unequal mean abilities affect only slightly the error rate. All methods have a slightly higher error rate when the mean abilities are different. The only exceptions are found for the robust methods when a normal distribution is used. Third, the two robust methods perform about equally well, but except for two out of the eight cases, the RIP followed by the robust regression (RIP + rob) does slightly

better than the equivalent method with logits of proportion of success (Logit(p) + rob). The differences are so small that it does not really pay off to do the modeling. It seems sufficient to use marginal proportions, and hence a nonparametric approach. This result implies that it is worth investigating whether not the delta plot method performs quite well if used with robust regression. Overall, the robust methods do clearly better than the three traditional methods.

7.3. Application to the Verbal Aggression Data

Both, lmer and WinBUGS are used to estimate the RIP model with the following identification restrictions: the mean ability in the two groups is zero, and the variance is equal. When unequal variances are allowed, the goodness of fit is not really better. No restrictions are imposed on the distributional parameters of the difficulties so that the group level difference will appear in the differences between the difficulty means. The results of both estimation methods agree very much. The deviance, AIC, and BIC of the lmer result for the RIP model are 8179, 8191, 8232, respectively. The model does slightly better than the RPRI model (see Table 2), except for the BIC. This means that there is no strong evidence for two different difficulty profiles, and that the degree of DIF as identified through the RIP model is rather minor.

This result is confirmed by the correlation of 0.92 (0.05) between the two profiles. The variance estimate of the abilities is 1.90 (0.19), and the variance of the difficulties are 1.31 (0.56) for women and also 1.31 (0.60) for men. The scatter plot is given in Fig. 1.

When the five methods are applied to the verbal aggression data set, the items identified as DIF depend on the method that is used, as shown in Table 6. The robust methods seem to flag fewer items as DIF than the traditional methods. This result can be explained as follows. First, the DIF seems somewhat asymmetrical, which may lead to DIF inflation when not a robust method is used. Second, perhaps DIF is actually gradual, and the bivariate RIP model may in fact be the appropriate one, so that there not really outliers and, therefore, the robust methods may flag fewer items as DIF.

One can see in the scatter plot of Fig. 1 that the items identified earlier as possible DIF items are the ones that deviate most from the 45° line: 6 and 12 are found above the line, and 14, 16, 17, 19, and 20 are found under the line. The different methods clearly converge in practice,

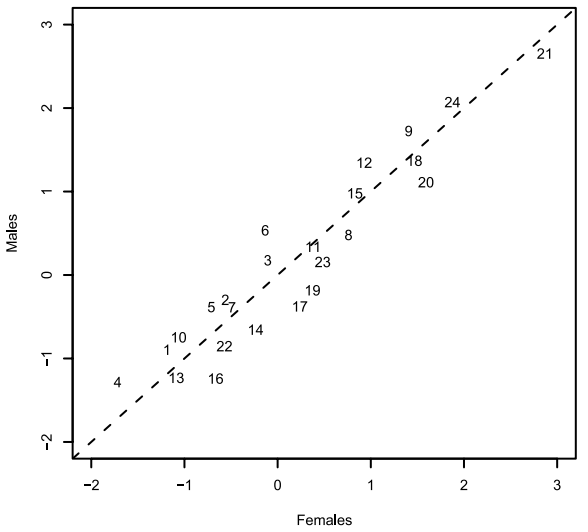

FIGURE 1.
Scatter plot of item difficulty estimates as obtained from the lmer estimation of the RIP model for the verbal aggression data.

TABLE 6.
Identification of DIF items in the verbal aggression data following the five methods.

Item # and description in terms of design factors	DIF sign[a]	MH	2 Rob methods	STD-P-DIF	LR
6: other to blame (train), want, shout	Positive	DIF	DIF[b]	DIF	DIF
12: self to blame (call), want, shout	Positive	DIF	–	DIF	DIF
14: other to blame (bus), do, curse	Negative	–	–	–	DIF
16: other to blame (train), do, curse	Negative	DIF	DIF	DIF	DIF
17: other to blame (train), do, scold	Negative	DIF	–	DIF	DIF
19: self to blame (grocery), do, curse	Negative	DIF	–	DIF	DIF
20: self to blame (grocery), do, scold	Negative	DIF	–	DIF	DIF

[a]Positive DIF value means that the item is "easier" for women, and negative DIF means that the item is "easier" for men.
[b]DIF means the item is flagged as a DIF item, and – means it is not.

although they are conceptually quite different. Looking at Table 6, it is interesting to see that given an equal degree of propensity to verbal aggression, men are more inclined to actually curse and scold (items 14, 16, 17, 19, 20), and hence to show blaming behavior, whereas women seem to have a stronger desire to shout when frustrated (items 6, 12), and hence to give expression to their frustration, rather than showing blaming behavior.

7.4. Anchoring Methods and the Random Item Mixture (RIM) Model

Most DIF detection methods imply a kind of anchoring in an implicit or explicit way. The reason for describing the anchoring methods here next is that the random item mixture (RIM) model can be seen as a posterior way of anchoring, while the existing anchoring methods are all based on an a priori anchor set.

In terms of the Rasch model, DIF is a difference in difficulty, but in order to determine a difference in difficulty, the scales of both groups needs to be aligned through linking. IRT scales have no fixed origin, so that their location is a matter of choice for the scales of both groups. A not so nice consequence is that whether or not an item shows DIF, depends on the scale linking, or in other words, on the anchoring method. Therefore, DIF is a relative concept, relative to the kind of linking that is chosen. In order to explain the linking issue and to prepare for the proposed solution, five anchoring methods will be described in the following, partly based on Wang (2004):

Equal mean ability anchoring means that the mean ability in both groups is set to a given common value, for example, 0.00:

$$\mu_{\theta_0} = \mu_{\theta_1} = 0. \tag{5}$$

Let us denote the original difficulties as β_{ig}^*, and the difficulties after linking β_{ig}, the difficulty of item i in group g ($g = 0$ for reference group, and $g = 1$ for focal group). The resulting DIF value for item i is $\delta_i = \beta_{i1} - \beta_{i0}$. Although this kind of anchoring solves the scale linking problem of the model, it is clearly not a good kind of anchoring because the difference in difficulty expresses also overall differences in ability between the two groups.

Equal mean difficulty anchoring means that the mean difficulty in both groups is set to a given common value, for example, 0.00. This corresponds also with the earlier mentioned ANOVA kind of linking. The anchoring requires a transformation of difficulties by subtracting

the mean $\beta^*_{.g}$, so that

$$\sum_{i=1}^{I} \beta_{ig} = 0, \tag{6}$$

where $\beta_{ig} = \beta^*_{ig} - \beta^*_{.g}$. The resulting DIF value is $\delta_i = (\beta^*_{i1} - \beta^*_{i0}) - (\beta^*_{.1} - \beta^*_{.0})$.

This anchoring method does not suffer from a contamination with an overall group difference, and it still solves the linking problem. Although it is common practice to anchor in this way, this method contains the risk of "expanding" DIF to items where there is no DIF, and to "shrinking" the degree of DIF of the true DIF-tems, as illustrated earlier.

The three alternatives to these two all work with a set of anchoring items, A, which are items with equal difficulty in both groups:

$$\beta_{i1} = \beta_{i0} \quad \text{if } i \in A = \{i \mid a_i = 0\}, \tag{7}$$

where a_i is the anchor set indicator, $a_i = 0$ if item i belongs to the anchor set, and $a_i = 1$ otherwise. The three methods differ in how membership of the anchor set A is determined.

Single-item anchoring means that the anchoring set consists of only one item, item i', so that $\beta_{i1} = \beta^*_{i1} - (\beta^*_{i'1} - \beta^*_{i'0})$, and $\delta_i = (\beta^*_{i1} - \beta^*_{i0}) - (\beta^*_{i'1} - \beta^*_{i'0})$, without any further linking constraints. Also this model solves the linking problem, but it does so in an arbitrary way, by picking one single item and, therefore, also the definition of DIF is arbitrary because it depends on just one item.

The first three anchoring methods are all three appropriate for linking purposes, but only the second, equal mean difficulty anchoring is also appropriate for the identification of DIF. The following two methods are both appropriate as DIF identification methods, but they do more than linking. They constrain the model further than is required for scale linking.

Multiple-item anchoring means that one specifies in advance more than just one item which does not show DIF, and hence more than one item is a member of the anchor set A. This method works well if the anchoring assumption holds. It implies prior knowledge, based on theory, expert judgment, or on earlier studies.

All-other anchoring means that each item i in turn is considered a potential DIF item, while all other items are considered to define the set of anchor items. This assumption is made in an implicit or explicit way for a variety of DIF indices, which do not require any modeling, but do make model assumptions nevertheless. This is, for example, the case for the Mantel–Haenszel (MH) method (Holland & Thayer, 1988), the logistic regression (LogReg) method (Swaminathan & Rogers, 1990), the standardization method (STD P-DIF) (Dorans and Kulick, 1986), and the Likelihood Ratio (LR) method (Thissen et al., 1986). All-other anchoring is used both for DIF identification, and as part of a heuristic. Following the heuristic procedure, an item identified as showing DIF is omitted, and the procedure is repeated with the remaining items. It is clear that the end result may depend on the order of the items as used in the heuristic.

It can be concluded from this short overview of anchoring methods that either assumptions need to be made (methods 4 and 5) which may lead to a misspecified model and, therefore, to misidentification of DIF, or the type of anchoring can yield misleading results: DIF confounded with the overall group difference (method 1), DIF inflation and DIF shrinkage in case of asymmetrical DIF (method 2), or arbitrary DIF (method 3). The best choice is the equal mean difficulty method (method 2) because it is does not yield a confounding with overall group differences, and because it doesn't hinge on an arbitrarily chosen item, while it also does not constrain the model in ways that might be incorrect.

The ideal method would be one with "posterior anchoring," which would mean that the anchor set A is a latent category, and that it is an issue of estimation to find out which items do

belong to A and which do not. This can be realized with a mixture model approach replacing the anchor set indicator a_i with a latent indicator α_i:

$$\beta_{i1} = \beta_{i0} \quad \text{if } i \in A = \{i \mid \alpha_i = 0\}, \tag{8}$$

where $\alpha_i = 0$ if item i is a non-DIF item, and $\alpha_i = 1$ if item i is a DIF item indeed.

Note that the latent anchor set is meant to be equivalent with the set of non-DIF items, whereas the manifest anchor set of the prior anchor method 4 (multiple-item anchoring) can be a subset of these. This posterior anchoring method leads to the model formulated in (3b).

The posterior approach has two important features. First, it fits with the discrete approach of DIF, implying that an item shows DIF or not, but it does so without distorting the results in case DIF is, in fact, gradual and can be captured by a RIP model. This is because the RIM model is a special case of the RIP model, with $\pi_\alpha = 1$ (all items being DIF items), to be explained when the model is described further. Second, the linking issue is solved based on the data and the estimation of the model, and not on an a priori choice as when one of the five anchoring methods is used. It makes DIF less relative because not the kind of a priori linking, but the data instead decide on the linking of the two scales.

7.5. Application of the RIM Model to the Verbal Aggression Data

The RIM model is a combination of the RIP model with an item mixture distribution:

$$\eta_{pig}|\alpha_i, \theta_{pg}, \beta_{i0}, \beta_{i1} = \theta_{pg} - (1 - \alpha_i g_p)\beta_{i0} - \alpha_i g_p \beta_{i1} + g_p \gamma, \tag{9}$$

where $\eta_{pig}|\alpha_i, \theta_{pg}, \beta_{i0}, \beta_{i1}$ is the logit of $\Pr(Y_{pig} = 1|\alpha_i, \theta_{pg}, \beta_{i0}, \beta_{i1})$, α_i is the latent binary item variable indicating whether item i is a DIF item, $\alpha_i \sim \text{Bernoulli}(\pi_\alpha)$, g_p is the group membership of person p: $g_p = 0$ if p belongs to the reference group, and $g_p = 1$ if p belongs to the focal group, where γ is the group difference parameter. θ_{pg} is the ability of person p belonging to group g, $\theta_{pg} \sim \text{N}(\mu_{\theta_g}, \sigma^2_{\theta_g})$ with μ_{θ_g} and $\sigma^2_{\theta_g}$ as the mean and variance of the ability in group g, and with identification restrictions $\mu_{\theta_0} = 0$ and $\sigma^2_{\theta_0} = \sigma^2_{\theta_1}$ (this latter restriction is not necessary for identification, but it was found not having a substantial effect either), β_{i0} is the difficulty of item i in the reference group, and β_{i1} is the difficulty of item i in the focal group if $\alpha_i = 1$, whereas the difficulty of item i in the focal group is β_{i0} if $\alpha_i = 0$, $(\beta_{i0}, \beta_{i1}) \sim \text{BVN}(\mu_{\beta_0}, \mu_{\beta_1}, \Sigma_{\beta_0\beta_1})$, with μ_{β_0} and μ_{β_1} as the mean difficulty in the reference group and also in the focal group if $\alpha_i = 0$, and the mean difficulty in the focal group if $\alpha_i = 1$, respectively.

For reasons of identification, the restriction $\mu_{\beta_0} = \mu_{\beta_1}$ is introduced. Without this equality, the model would not be identified because γ and μ_{β_1} can compensate for one another. This does not mean, however, that the RIM imposes DIF to be symmetrical. What matters for determining whether DIF is symmetrical or not, is the mean difficulty for the items with a high posterior probability of belonging to the DIF class.

Two special cases of the RIM model are of interest to see the potential of the model. The first special case is when the mixing probability $\pi_\alpha = 0$, which means that there is no DIF, and that the RIM reduces to the simple RPRI Rasch model. The second special case is when the mixing probability $\pi_\alpha = 1$, which means that all items show DIF in the focal group, so that the RIM model reduces to the RIP model. Hence, one may use the estimate of π_α as a diagnostic:

iff $\pi_\alpha = 0$, then RPRI Rasch;
iff $0 < \pi_\alpha < 1$, then RIM;
iff $\pi_\alpha = 1$, then RIP.

WinBUGS is used to estimate the model. The estimates are as follows: $\gamma = 0.27$ (0.23), $\sigma^2_\theta = 1.95$ (0.19), $\sigma^2_{\beta_0} = 1.32$ (0.38), $\sigma^2_{\beta_1} = 1.31$ (0.47), $r_{\beta_0\beta_1} = 0.80$ (0.18), $\pi_\alpha = 0.71$ (0.21).

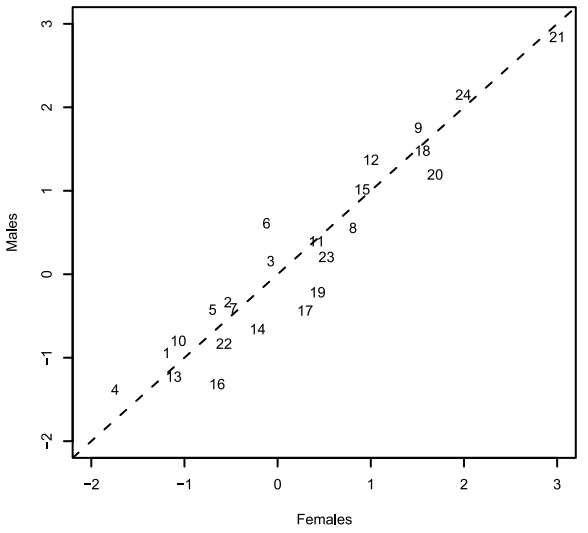

FIGURE 2.
Scatter plot of item difficulty estimates as obtained from the lmer estimation of the RIM model for the verbal aggression data.

There is no significant overall difference between the two groups. When the difficulties for the two item mixture components are plotted as in Fig. 2, the resulting figure is very similar to the one in Fig. 1. The RIM model does not coincide with the RIP model in this application because the estimate for π_α is smaller than one ($\pi_\alpha = 0.71$), but on the other hand, the posterior probabilities of all items are larger than 0.50 (see Fig. 3). It is interesting to see how the items 6, 12, 14, 16, 17, 19, and 20 have the highest posterior probabilities. In sum, this RIM model is hardly differentiated from the RIP model. Hence, the results obtained with the RIM model are not of the kind that a discrete kind of DIF is supported. Rather, DIF seems of the RIP type, given the high value of the mixing probability estimate and all items being identified as DIF items, based on their posterior probability.

Unfortunately, this also means that the data set is not optimal to illustrate the potential of the RIM model for DIF identification. Therefore, we decided to generate new data for the items 14, 16, 17, 19, 20, and 23 (with the largest distances from below the bisector in Fig. 1), following the RIP model, but with difficulties for males that are either 1.00 or 2.00 lower than the estimated values based on the original data. The corresponding newly generated data sets are called verbal aggression data minus one (VA-1) and minus two (VA-2), to differentiate them from the original VA data.

The RIM model was again estimated with WinBUGS. The corresponding mixing probabilities are 0.476 (0.159) and 0.413 (0.123) for the VA-1 and VA-2 data, respectively. Figures 4 and 5 give the corresponding posterior probabilities of belonging to the DIF class if the RIM model is used.

It is clear from Figs. 4 and 5 that the situation is now quite different than for the original VA data. The mixing probability estimates are much lower than for the original VA-data, but still substantially higher than zero. The posterior probabilities shown in Figs. 4 and 5, do clearly indicate the items 14, 16, 17, 19, 20, and 23 now as DIF items (as it should be, based on the data manipulation), but in Fig. 4 (VA-1 data) also items 6, 8, and 22 have a posterior probability higher than 0.50, although clearly lower than the posterior probability of the six. This is not surprising given the position of the items 6, 8, and 22 in Fig. 1. Item 6 is located above the bisector and items 8 and 22 are located below the bisector. The situation depicted in Fig. 5 (VA-2 data) is

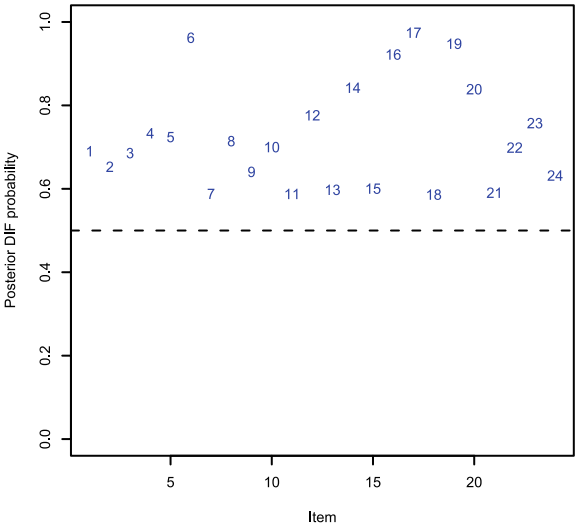

FIGURE 3.
Posterior probabilities of belonging to the DIF class of items.

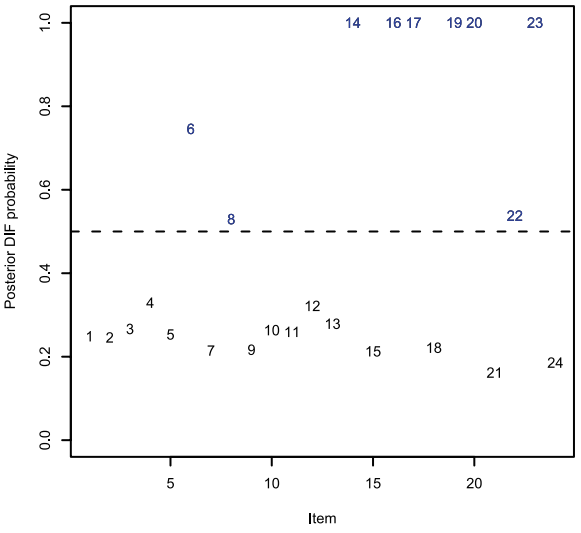

FIGURE 4.
Posterior probabilities of belonging to the DIF class of items for the VA-1 data.

very similar, but now with decreased posterior probabilities, so that only item 6 exceeds the 0.50 threshold.

These results illustrate that the RIM model succeeds in correctly flagging the DIF items indeed, and that it avoids making false identifications. The items 6, 8, and 22, and especially item 6, may perhaps be considered DIF items, although not involved in the data manipulation. Note that DIF is asymmetrical, and that an equal mean difficulty anchoring would export DIF to all items, while in this application, a quite good identification of DIF items was realized. Frederickx, Tuerlinck, De Boeck, and Magis (2008) show through several simulation studies that in fact the RIM model and the resulting posterior probabilities do a better job identifying DIF

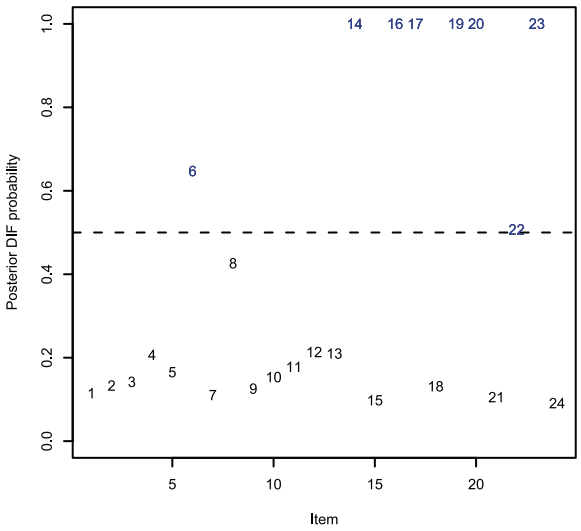

FIGURE 5.
Posterior probabilities of belonging to the DIF class of items for the VA-2 data.

items than the three traditional methods investigated earlier. This suggests that the RIM model is a successful tool for the identification of DIF.

7.6. Discussion

Two concepts have been developed as an alternative for the common practice. One is the RIP concept of DIF, where DIF is no longer a discrete event, but described instead with a bivariate or multivariate distribution of item difficulties. The other is the RIM concept, adding a discrete but random element to the previous, by introducing a latent binary item variable to indicate DIF. Both concepts share that they do not need any kind of a priori anchoring and neither are they iterative. They differ with respect to what can (or should) be done when DIF is unwanted. Following the RIP concept, there is no way one can adapt the test to eliminate DIF because DIF is actually bidimensionality in the item difficulties, introduced by the inclusion of different kinds of persons, just as the bidimensionality of individual differences is introduced by the inclusion in the test of different kinds of items. However, in a next step, one can use a procedure such as robust regression in order to flag DIF items as illustrated earlier. Following the RIM concept, one can adapt the test, omitting all items with a posterior probability of 0.50 of higher of belonging to the DIF class.

Both the RIP and the RIM approach require modeling, and are therefore vulnerable to misspecifications of the model, whereas the robust regression approach can either start from model estimation results, or can operate with simple marginal statistics without any modeling. While the robust regression approach can be used for a discrete DIF concept as well as for a gradual DIF concept, it does lead to DIF identification anyhow. The quality of the identification seems equally good as that of other methods if DIF is symmetrical, and better if DIF is asymmetrical. A limitation of the method is that it requires DIF to be a minority phenomenon among the items. On the other hand, one may wonder whether DIF does not lose its meaning if the majority of items shows DIF, because then it is no longer a disturbance, but the major phenomenon, so that it must be concluded that the difference between the groups is really of a qualitative kind (De Boeck, Wilson, & Acton, 2005), and that the items tap different kinds of abilities in the two groups. The same issue can be raised for random item profiles in a RIP model when their correlation is near zero. Can one still consider the abilities as being similar?

It would require further simulation studies, for example, both with a larger proportion and a smaller proportion of DIF items, with a larger range of DIF values, etc. in order to assess the qualities of the three approaches illustrated here (robust approach, RIP, RIM) in comparison with the more traditional methods.

8. General Discussion and Conclusion

Random item models are a rather new approach in the domain of IRT. Because they fit into a generalizability concept, they have a clear conceptual asset. Ideally, one wants the measurement of a person to generalize over a domain of similar items, instead of being restricted to the particular item set under consideration (Briggs & Wilson, 2007). Random item models have also some interesting assets from a statistical point of view. The number of parameters does not proliferate, but remains limited instead to distributional parameters, and the random item models fit in with the general way of using error components to deal with imperfect relations.

On the other hand, random item models also have some drawbacks. First, when also the persons are random, then the model becomes a crossed random effects model, which is a complication for the estimation, and the available estimation algorithms are limited in number and experience one has with using them. Second, the notion of random items is still controversial, because it is not always clear what the population of items would be, and what it would mean for items to be drawn from that population. Third, the number of items in a test is rather limited, so that the information for model estimation is limited as well and, therefore, its basis is rather narrow.

The illustration of random item models was restricted to Rasch type of models. It is a challenge to extend the approach to the 2PL type of models. However, this also opens new perspectives, such as random item discriminations. Often the estimation of the discrimination parameters is not very robust. Random discrimination models may contribute to the stability of the 2PL model. This is a topic that deserves further investigation.

In sum, it may be concluded from the application of random item models to the verbal aggression data set, that these models are promising and useful to solve various kinds of issues. Of course, further investigation is required for a more definite conclusion.

Acknowledgements

The author wants to thank David Magis for his invaluable help with the analyses, and Francis Tuerlinckx for his advice. The research reported here was supported by the Fund for Scientific Research (FWO) grant G.0148.04 and by the K.U. Leuven Research Council grant GOA/2005/04.

References

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

Albers, W., Does, R.J.M.M., Ombos, Tj., & Janssen, M.P.E. (1989). A stochastic growth model applied to tests of academic knowledge. *Psychometrika*, *54*, 451–466.

Andersen, E.B. (1980). *Discrete statistical models with social science applications*. Amsterdam: North-Holland.

Angoff, W.H., & Ford, S.F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement*, *10*, 95–106.

Bates, D., Maechler, M., & Dai, B. (2008). The lme4 Package version 0.999375-26. http://cran.r-project.org/web/packages/lme4/lme4.pdf/.

Bejar, I.I. (1993). A generative approach to psychological and educational measurement. In N. Frederiksen, R.J. Mislevy, & I.I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 323–359).

Bejar, I.I., Lawless, R.R., Morley, M.E., Wagner, M.E., Bennett, R.E., & Revuelta, J. (2003). A feasibility study of on-the-fly item generation in adaptive testing. *Journal of Technology, Learning, and Assessment*, *2*, 1–29.

Bock, R.D., & Mislevy, R.J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*, 431–444.

Briggs, D.C., & Wilson, M. (2007). Generalizability in item response modeling. *Journal of Educational Measurement*, *44*, 131–155.

Camilli, G., & Shepard, L.A. (1994). *Methods for identifying biased test items*. Sage: Thousand Oaks.

Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in proficiency tests. *Language Testing*, *2*, 155–163.

Cho, S.-J., & Rabe-Hesketh, S. (2008). *Estimating item response models with random item parameters*. Unpublished manuscript.

Clark, H.H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*, 335–359.

Coleman, E.B. (1964). Generalizing to a language population. *Psychological Reports*, *14*, 219–226.

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

De Boeck, P., Wilson, M., & Acton, S. (2005). A conceptual and psychometric framework for distinguishing categories and dimensions. *Psychological Review*, *112*, 129–158.

Dorans, N.J., & Holland, P.W. (1993). DIF detection and description: Mantal-Haenszel and standardization. In P.W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale: Erlbaum.

Dorans, N.J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, *23*, 355–368.

Embretson, S.E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika*, *64*, 407–433.

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

Frederickx, S., Tuerlinckx, F., De Boeck, P., & Magis, D. (2008). *An item mixture model to detect differential item functioning.* Unpublished manuscript, K.U. Leuven.

Gelman, A., & Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–511.

Glas, C.A.W., & van der Linden, W.J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement*, *27*, 247–261.

Hively, W., Patterson, H.L., & Page, S.H. (1968). A "universe-defined" system of arithmetic achievement tests. *Journal of Educational Measurement*, *5*, 275–290.

Holland, P.W., & Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & J.I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale: Lawrence Erlbaum.

Holland, P.W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale: Lawrence Erlbaum.

Ironson, G.H., Homan, S., Willis, R., & Singer, B. (1984). The validity of item bias techniques with math word problems. *Applied Psychological Measurement*, *8*, 391–396.

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

Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 189–212). New York: Springer.

Johnson, P.M., & Sinharay, S., (2005). Calibration of polytomous item families using Bayesian hierarchical modeling. *Applied Psychological Measurement*, *29*, 369–400.

Lunn, D., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS—a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719–748.

McGraw, K.O., & Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30–46.

Millsap, R.E., & Everson, H.T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*, 297–334.

Raaijmakers, J., Schrijnemakers, J., & Gremmen, F. (1999). How to deal with "the language-as-fixed-effect-fallacy": Common misconceptions and alternative solutions. *Journal of Memory and Language*, *41*, 416–426.

Popham, W.J. (1978). *Criterion-referenced measurement*. Englewood Cliffs: Prentice-Hall.

Rouder, J.N., Lu, J., Speckman, P.L., Sun, D., Morey, R.D., & Naveh-Benjamin, M. (2007). Signal detection models with random participant and random item effects. *Psychometrika*, *72*, 621–624.

Rousseeuw, P.J., & Leroy, A.M. (1987). *Robust regression and outlier detection*. New York: Wiley.

Rousseeuw, P.J., & van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, *41*, 212–223.

Roussos, L.A., Templin, J.L., & Henson, R.A. (2007). Skills diagnosis using IRT-based latent class models. *Journal of Educational Measurement*, *44*, 293–311.

Savalei, V. (2006). Logistic approximation to the normal: The KL rationale. *Psychometrika*, *71*, 763–767.

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlation: Uses in assessing reliability. *Psychological Bulletin*, *86*, 420–428.

Shepard, L., Camilli, G., & Williams, D.M. (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement*, *22*, 77–105.

Sinharay, S., Johnson, M.S., & Williamson, D.M. (2003). Calibrating item families and summarizing the results using family expected response functions. *Journal of Educational and Behavioral Sciences*, *28*, 295–313.
Snijders, T.A.B., & Bosker, R.J. (1999). *Multilevel analysis. An introduction to basic and advanced multilevel modeling*. London: Sage.
StataCorp (2007). *Stata statistical software: Release 10*. College Station: StataCorp LP.
Swaminathan, H., & Rogers, H.J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*, 361–370.
Tan, E.S., Ambergen, A.W., Does, R.J.M.M., & Imbos, Tj. (1999). Approximations of normal IRT models for change. *Journal of Educational and Behavioral Statistics*, *24*, 208–223.
Teresi, J.A. (2001). Statistical methods for examination of differential item functioning (DIF)—with applications to cross-cultural measurement of functional, physical and mental health. *Journal of Mental Health and Aging*, *7*, 31–40.
Thierny, L., & Kadane, J.R. (1986). Accurate approximations for the posterior moments and marginal densities. *Journal of the American Statistical Association*, *81*, 82–86.
Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin*, *99*, 118–128.
Tuerlinckx, F., Rijmen, F., Verbeke, G., & De Boeck, P. (2006). Statistical inference in generalized linear mixed models: A review. *British Journal of Mathematical and Statistical Psychology*, *59*, 225–255.
Van den Noortgate, W., De Boeck, P., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics*, *28*, 369–386.
Verhelst, N.D., & Eggen, T.J.H.M. (1989). *Psychometrische en statistische aspecten van peilingsonderzoek* (PPON rapport 4). Arnhem: Cito.
Wang, W.-C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *Journal of Experimental Education*, *72*, 221–261.
Zwinderman, A.H. (1991). A generalized Rasch model for manifest predictors. *Psychometrika*, *56*, 589–600.

Published Online Date: 2 DEC 2008