Psicológica (2014), 35, 309-337.

Fixed item parameter calibration for assessing differential item functioning in computerized adaptive tests

González-Betanzos, F.*[1], Abad, F. J. **, Barrada, J. R.***

**Universidad Michoacana de San Nicolas de Hidalgo, Morelia, México*

***Universidad Autónoma de Madrid, Madrid, Spain*

****Universidad de Zaragoza, Teruel, Spain*

In computerized adaptive testing pretest items are presented in conjunction with operational items to renew the item bank. Pretest items are calibrated, and possible differential item functioning (DIF) is analyzed. Some difficulties arise due to the large amount of missing responses, which can be avoided by the use of fixed item parameter calibration (FIPC; Kim, 2006) methods. In this study, we applied the multiple weights updating and multiple EM cycles method, with response imputation (as suggested by Lei, Chen, & Yu, 2006) and without response imputation for non-applied items. The IRT likelihood ratio test (IRT-LRT) was used for DIF detection. The manipulated factors were type of DIF, DIF size, impact size, test length, and sample size. The results showed that the FIPC method is suitable for detecting large-size DIF in large samples. In the presence of impact the use of imputation led to a bias in the effect-size measure of the DIF.

In computerized adaptive testing (CAT) programs, periodic replacement of obsolete or overexposed items is required (Mills & Stocking, 1996). For that purpose, new (pretest) items are usually presented to examinees during the course of their testing with already calibrated and

[1] This research was partly supported by a grant from the Spanish Ministry of Education and Science [PSI2009-10341] and by the UAM-IIC Chair Psychometric Models and Applications; The first author was supported by a doctoral fellowship from the PROMEP program of the Mexican Secretary of Public Education [UMSNH-EXB-205]. Author's address: Fabiola González-Betanzos**.** Facultad de Psicología, UMSHN**.** 58120-Mich., México. Telephone: 00 52 4433145987. E-mail: betanzos@umich.mx

in-use (operational) items. In the simplest design, analyzed in the current study, a fixed (nonadaptive) small block of pretest items –even a single item–, not considered for scoring test takers, is embedded within the CAT. One advantage of the online calibration design is that pretest items are presented to representative and motivated samples in a real evaluation situation (Parshall, 1998). Subsequently, pretest items can be calibrated and must be routinely checked for differential item functioning (DIF).

DIF occurs when the conditional probability of a correct response, given the latent ability, differs between two compared groups (termed the reference and the focal group). Detecting DIF is important because DIF can invalidate procedures for making decisions about individuals. DIF items should be especially avoided in CAT because fewer items are administered to estimate the examinees' test scores, and the sequence of the administered items may be influenced by the responses to the flawed items (Zwick, 2010). Furthermore, in fixed tests, the item DIF may cancel each other out, resulting in minimal test bias at the score level, whereas in CATs, the DIF will always depend on the set of items administered (Steinberg, Thissen, & Wainer, 1990).

Pretest on-line calibration design is being use increasingly to renew the item pool (Pommerich & Segall, 2004). However, some issues have been identified in the implementation of DIF methods in the CAT environment (Lei, Chen, & Yu, 2006; Nandakumar & Roussos, 2004; Zwick, 2010). One primarily affects DIF methods where examinees are matched by their direct scores (sum of correct items), which is no longer appropriate in CATs (Harmes, Parshall, & Kromrey, 2003; Steinberg et al., 1990), where the number of correct responses is expected to be roughly the same for all the examinees with independence of their trait level. Consequently, some DIF methods such as Mantel-Haenzel standardization, SIBTEST and logistic regression have been adapted by using CAT ability estimate (Lei et al., 2006; Nandakumar & Roussos, 2004) or expected score over the entire item pool based on CAT ability estimate (Zwick, Thayer, & Wingersky, 1993, 1994a, 1994b, 1995) as matching variables.

Additional problems for DIF detection in CATs arise from the incompleteness of the response matrix, as there will be a huge proportion of missing responses for adaptively applied operational items and many items may be responded by very few examinees. Also, operational item responses are based on a restricted range of ability. If focal and reference group distributions for the latent trait differs, there should be little overlap in the anchor operational items.

Because item response theory (IRT) is required in CAT development, the IRT likelihood ratio test (IRT-LRT) appears to be a logical choice for assessing DIF in the CAT environment (Miller, 1992). Thissen, Steinberg, and Wainer (1988, 1993) proposed the IRT-LRT, in which the log-likelihood is compared between two different models (one in which item parameters are set equally in the focal and the reference group and another in which these parameters are allowed to vary). With IRT-LRT, both parameter estimates of pretest items and DIF analysis are obtained at the same time. IRT-LRT technique offers the additional advantage of being applicable directly to CAT data. Unlike other DIF methods, the IRT-LRT does not rely on the direct score. In applying this IRT procedure in the online calibration scenario, we need to calibrate the pretest items along with the operational items for each group (reference and focal); if the test indicates DIF presence then a DIF effect size should be computed. There are two alternative strategies to calibrate pretest items: concurrent calibration with linking and fixed item parameter calibration (Kim, 2006).

In the concurrent calibration with linking method, both operational and pretest item parameters are estimated together at the first step without considering the existing scale for operational items. Then, in a second step, pretest item parameters are placed onto the existing scale by equating using the operational items as linking items (Ban, Hanson, Wang, Yi, & Harris, 2001; Hsu, Thompson, & Chen, 1998; Stocking, 1988). This leads two main problems. First, CAT data are not optimal for operational item parameter estimation due to data sparseness of the response matrix and a restricted range of ability of the examinees responding to the items (Haynie & Way, 1995). For example, some operational items can be rarely applied and thus sample size for its calibration will be very small. Furthermore, lack of item overlap due to the adaptive application increase the problem of sparseness. Harmes, Parshall, and Kromrey (2001) found that calibration software (v.g., BILOG) was unable to successfully calibrate sparse response data matrices without the application of missing data treatments. Second, the pretest item parameters should be transformed onto the established (operational) scale by a linking step that might add an error to the item parameters and group distribution estimates (Ban et al., 2001). This problem can be especially severe as operational item parameters are being estimated within a problematic design.

In contrast, fixed item parameter calibration (FIPC) does not involve a linking step. In FIPC the operational item parameters are fixed to their previously estimated parameter values and only the pretest items are calibrated (Kim, 2006). FIPC has been found to be successful in calibrating

pretest items in the CAT environment (Ban et al., 2001). Up to now, FIPC has not been proposed to assess DIF with CATs.

At present, only one study (Lei et al., 2006) has used the IRT-LRT for DIF analysis in CAT. Lei et al. (2006) suggested that the IRT-LRT has not been examined in the CAT environment due largely to the complexity of estimation when there is a huge proportion of non-randomly missing data in the operational items. To overcome that issue, Lei et al. suggested "[t]o capitalize, on the advantages of IRT, by imputing data from the assumed model for the empty cells" (p. 247). The imputed responses were generated based on the probabilities of getting the items correct, computed from the three-parameter logistic model using the known item parameters and CAT ability estimates. Results for IRT-LRT with imputation were promising. However, with this DIF detection design, equating is still required for recovering pretest item parameters in the adequate metric for the concurrent calibration method. And, importantly, CAT ability estimates are used for imputation, but estimation uncertainty is ignored (Harmes et al., 2003; Little & Rubin, 1987).

Thus, the only study to date that has evaluated the appropriateness of the IRT-LRT method for detecting DIF on CATs (Lei et al., 2006) used a concurrent calibration method and imputed a large proportion of the data set. Despite the general satisfactory performance, Lei at al. (2006) acknowledged that IRT-LRT approach used to DIF detection remained questionable. We considered that in order to ensure the smooth transition to CATs data, FIPC without imputation may be a better-suited and efficient alternative for detecting DIF on pretest items than concurrent calibration. Moreover, if operational item parameters are fixed and IRT standard assumptions hold, imputation would not be necessary and in fact should be avoided because it might introduce distortions on the item parameters and group distribution scale if the CAT ability has been estimated with high error levels.

This study analyze the performance of the IRT-LRT in detecting DIF on pretest items seeded in CATs, using FIPC to compare two ways of handling the missing data problem, with imputation (as in Lei et al., 2006) and without imputation. We will first describe the IRT-LRT and FIPC in greater detail.

Item Response Theory Likelihood Ratio Test in the CAT Environment

Two different models are estimated for the assessment of DIF. In the first one, called the compact model, the parameters of all the items for the

reference and the focal group are constrained to be equal. In the second one, the augmented model, the parameters of the anchor items are fixed equally across the groups, and all (or some) of the parameters of the item tested for DIF vary. The DIF will be detected when the fit of the augmented model is significantly better than that of the compact model. The log-likelihood of the models is the measure of the fit, and the log-likelihood ratio (LR) is a measure of the fit increment for the augmented model:

$$LR = -2\ln L_C - (-2\ln L_A), \quad (1)$$

where L_C is the log-likelihood of the compact model, and L_A is the log-likelihood of the augmented model. The IRT-LR is approximately χ^2 distributed, with degrees of freedom equal to the difference in the number of free parameters. If the general test is significant, an omnibus test is easily carried out to establish whether DIF is due to differences in *a*, *b* or *c* parameters when a 3PL is used. These tests are conducted by constraining the individual item parameters in the compact model. Because power and Type I error rates for the follow-up test are highly dependent on those for the general test, only the general test is evaluated in the present study.

IRTLRDIF (Thissen, 2001) is the software usually used to perform the estimations required for this technique of DIF detection. Thissen (2001) cited DIF detection in CAT as an important motivation for the development of IRTLRDIF (p. 14). This was the software used in the study of Lei et al. (2006), in which one subset of the operational items was selected as anchor items, and the pretest items were tested for DIF. IRTLRDIF performs concurrent calibration. Thus, it estimates all the item parameters, including those from the operational items (already known) and those from the pretest item (those unknown and of interest). This implies unnecessary computational time and possible convergence problems. More importantly, with IRTLRDIF an additional step is required to place the pretest item parameters onto the operational established scale (Kim, 2006).

Lei et al. (2006) evaluated the proposed IRTLRDIF adaptation method. They compared IRTLRDIF with the adapted SIBTEST, called CATSIB (Nandakumar & Roussos, 2004), and the logistic regression (LR) adapted by replacing the total score with the CAT trait estimate. They found that only the IRT-LRT provides adequate Type I error control consistently across all the sample-size and distribution conditions. In LR, Type I inflation was found to occur under impact conditions, which the authors partially attributed to the statistical bias introduced by the impact. In CATSIB, the inflation was noted to occur under impact and unequal

sample-size conditions, which the authors attributed to the low frequencies for some subgroups.

Furthermore, the statistical power in detecting DIF was generally higher with a higher DIF size effect, under no impact, and with equal sample-size conditions. The IRT-LRT and LR were found to be more powerful than CATSIB in detecting non-unidirectional DIF, whereas CATSIB was more powerful in detecting unidirectional DIF (especially when the DIF size effect was small). Among the two methods considered to be suitable for non-unidirectional DIF detection in non-impact conditions, the IRT-LRT generally outperformed LR under unequal sample-size conditions, whereas LR was more powerful with equal sample-size conditions. The IRT-LRT was found to have lower power mainly in detecting difficult and low-discriminating non-unidirectional DIF items.

Despite the generally satisfactory performance of the IRT-LRT in detecting DIF, the above-noted computational cost limits its use. In addition, not all the operational items could be treated as anchor items because IRTLRDIF was unable to handle so many items. Instead of treating 300 items as anchors (the whole bank), Lei et al. (2006) used only 84 (those from a single content area). Problems partially arise due to the requirement of recalibrating operational items and thus can be avoided by the use of fixed parameter calibration (FIPC) methods. In FIPC, operational item parameters are fixed to their previously estimated values, and only pretest items are calibrated (Ban et al., 2001; Ban, Hanson, Yi, & Harris, 2002; Kim, 2006). We will now show how FIPC could be used to assess DIF in CAT environments.

FIPC Methods Applied to DIF Detection

Suppose that each examinee i of group g (g = 1 {reference}, 2 {focal}), denoted by i_g (i_g = 1, ..., N_g, where N_g is the number of examinees in the group g), respond to one adaptive test, in which the items are selected from a set of $J_{(ope)}$ previously calibrated operational items and to $J_{(pre)}$ pretest items whose parameters are unknown. The likelihood of responses for each i_g examinee, given the set of applied items (S), would be

$$f(\mathbf{u}_{i_g} \left| \theta_{i_g}, \mathbf{\Delta}_g \right.) = \prod_{\substack{j=1 \\ j \in S}}^{J} P_j(u_{i_g j} \mid \theta_{i_g}, \delta_{jg}), \qquad (2)$$

where J is the number of applied items ($J = J_{(ope)} + J_{(pre)}$), $\mathbf{u}_{i_g}$ is the vector with the item observed responses ($u_{i_g j} = 0,1$), θ_{i_g} is the latent trait for

examinee i_g, and $\mathbf{\Delta_g} = (\boldsymbol{\delta}_{1g}...\boldsymbol{\delta}_{Jg})$ denotes the collection of item parameters for the group g, where $\boldsymbol{\delta}_{jg}$ is the vector of parameters for item j in group g. The complete sets of examinees and items, as well as of parameters, is denoted as $\boldsymbol{\theta}$ and $\mathbf{\Delta} = (\mathbf{\Delta}_1, \mathbf{\Delta}_2)$, respectively. Another important distinction in an online calibration setting is between the pretest and operational item parameters $\{\mathbf{\Delta_g} = (\mathbf{\Delta}_{g(pre)}, \mathbf{\Delta}_{g(ope)})\}$.

The EM algorithm may be used to find the values of the pretest item parameters, $\mathbf{\Delta}_{(pre)}$, when $\mathbf{\Delta}$ and $\boldsymbol{\theta}$ are simultaneously unknown. In each EM cycle l, E (expected) and M (maximization) steps are performed. In the E step, the posterior probabilities for θ_{k_g} at each of K_g quadrature points ($k_g = 1,2,...K_i$) are estimated assuming the provisional item and distribution parameters estimated in the previous cycle, $\hat{\mathbf{\Delta}}_g^{(l-1)}$ and $\hat{\boldsymbol{\pi}}_g^{(l-1)}$:

$$f(\theta_{k_g} \mid \mathbf{u}_{i_g}, \hat{\mathbf{\Delta}}_g^{(l-1)}, \hat{\boldsymbol{\pi}}_g^{(l-1)}) = \frac{f(\mathbf{u}_{i_g} \mid \theta_{k_g}, \hat{\mathbf{\Delta}}_g^{(l-1)}) f(\theta_{k_g} \mid \hat{\boldsymbol{\pi}}_g^{(l-1)})}{\sum_{k_g=1}^{K_g} f(\mathbf{u}_{i_g} \mid \theta_{k'_g}, \hat{\mathbf{\Delta}}_g^{(l-1)}) f(\theta_{k'_g} \mid \hat{\boldsymbol{\pi}}_g^{(l-1)})}, \quad (3)$$

where $\hat{\boldsymbol{\pi}}_g^{(l-1)}$ refers to the provisional distribution parameters that determine $f(\theta_{k_g} \mid \hat{\boldsymbol{\pi}}_g^{(l-1)})$, the prior weights of θ_{k_g} for group g. For example, if the group distributions are assumed to be normal, $\hat{\boldsymbol{\pi}}_g$ would contain $\hat{\mu}_g$ and $\hat{\sigma}_g$, the mean and standard deviation for group g, respectively, and $f(\theta_{k_g} \mid \hat{\boldsymbol{\pi}}_g)$ would be the prior normal distribution weight for the quadrature point θ_{k_g}. Provisional initial estimates, $\hat{\mathbf{\Delta}}^{(0)}$ and $\hat{\boldsymbol{\pi}}^{(0)}$, are used in the first E step.

The M step finds the $\hat{\mathbf{\Delta}}_g^{(l)}$ and $\hat{\boldsymbol{\pi}}_g^{(l)}$ parameters that maximize the conditional expectation of the complete data log-likelihood, in which the expectation is taken with respect to the conditional distribution of the missing data, given the observed data and some fixed known values of the parameters (Dempster, Laird, & Rubin, 1977; Woodruff & Hanson, 1996):

$$\left(\hat{\mathbf{\Delta}}^{(l)}, \hat{\boldsymbol{\pi}}^{(l)}\right) = \arg\max_{\mathbf{\Delta},\boldsymbol{\pi}} \sum_{g=1}^{2} \sum_{i_g=1}^{N_g} \sum_{k_g=1}^{K_g} \left\{ \log\left[f(\mathbf{u}_{i_g} \mid \theta_{k_g}, \mathbf{\Delta}_g) f(\theta_{k_g} \mid \boldsymbol{\pi}_g) \right] f(\theta_{k_g} \mid \mathbf{u}_{i_g}, \hat{\mathbf{\Delta}}_g^{(l-1)}, \hat{\boldsymbol{\pi}}_g^{(l-1)}) \right\}, \quad (4)$$

where $\hat{\mathbf{\Delta}}^{(l)}$ and $\hat{\boldsymbol{\pi}}^{(l)}$ $\{\hat{\boldsymbol{\pi}}^{(l)} = (\hat{\boldsymbol{\pi}}_1^{(l)}, \hat{\boldsymbol{\pi}}_2^{(l)})\}$ can be estimated separately.

A scale indeterminacy occurs if all the item parameters (operational and pretest) are estimated because the IRT ability scale is determined up to a linear transformation (e.g., Lord, 1980; pp. 36-38). Additionally, in DIF testing, the following constraints should be considered for model identification purposes: (a) the mean and standard deviation of the distribution of ability in the reference group are set at some specified values (e.g., $\mu_1 = 0$ and $\sigma_1 = 1$); and (b) the parameters of a set of items, the anchor items, are constrained to be equal across the groups (e.g., $\hat{\boldsymbol{\Delta}}_{1(ope)} = \hat{\boldsymbol{\Delta}}_{2(ope)}$). One problem with these constraints is that the parameter estimates ($\hat{\boldsymbol{\Delta}}_{(pre)}$ and $\hat{\boldsymbol{\pi}}_2$) may be obtained on a metric scale that differs from the original operational item metric scale (v.g., if the reference group in the calibration study have $\mu_1 \neq 0$ or $\sigma_1 \neq 1$). Furthermore, the recalibration of rarely applied operational items can be problematic.

In an FIPC method, only the pretest item parameters have to be estimated. Operational items are also used as anchor items, but their parameters are fixed to their known values. This avoids the recalibration of operational items and solves the scale indeterminacy problem without additional constraints, allowing $\hat{\boldsymbol{\pi}}_1$ estimates (v.g., μ_1 and σ_1 are not assumed to be zero and one, respectively). Another additional advantage is that the $\hat{\boldsymbol{\Delta}}_{(pre)}$ and $\hat{\boldsymbol{\pi}}_2$ parameters are obtained on the operational item metric scale.

Ban et al. (2001, 2002) and Kim (2006) suggested several variants in the application of the FIPC methods. Specifically, these FIPC methods are conceptually distinguished by the use of pretest item responses in estimating posterior probabilities that affect the number of times the prior ability distribution is updated and the number of EM cycles used. Kim (2006) compared these methods in terms of recovery of the trait distribution and pretest item parameters. In the said study, responses to a fixed test of operational and pretest items were simulated according to the 3PL model. First, the operational items were previously calibrated in a sample in which ability parameters were drawn from a standard normal distribution. Then, the operational item parameters were treated as fixed in the FIPC of pretest items. The results showed that only the multiple weights updating procedure (MWU-MEM), which uses operational and pretest item responses to update posterior probabilities at each EM cycle, produced proper results when the latent trait distribution in the new calibration sample differed from the standard normal distribution in the old calibration sample. Under these circumstances, the latent trait mean and variance parameter estimates had

larger biases in the remaining procedures. Specifically, these biases were larger when the number of operational items was smaller. Finally, Kim found that a relatively large sampling error in estimating the operational fixed item parameters with a small sample (i.e., 300 examinees) did not appear to severely affect the estimation of both the underlying ability distributions and the new item parameters. From these results, we selected the MWU-MEM as the potentially best FIPC method for DIF analyses in the presence of impact, in which the updating of prior weights, $\hat{\boldsymbol{\pi}}_1$ and $\hat{\boldsymbol{\pi}}_2$, is critical.

In summary, the primary purpose of this study is to evaluate the performance of the IRT-LRT obtained with FIPC in detecting the DIF of one pretest item seeded in a CAT. We propose FIPC as a more efficient method than the concurrent calibration because it does not require recalibrating operational items. Two methods were compared: FIPC with imputation and FIPC without imputation. Imputation has been claimed to be useful as a strategy to deal with missing responses to the operational items (Lei et al., 2006). However, single imputation does not consider uncertainty in the imputation procedure (Harmes et al., 2003; Little & Rubin, 1987). Accordingly, imputation is expected to produce worse results compared with non-imputation. This effect is expected to be larger when the CAT precision is lower (i.e., in a shorter CAT). Although we expect the application of the imputation-based method to be less successful, we include it because the only study that has applied the IRT-LRT to CATs advises its use (Lei et al., 2006).

METHOD

Conditions for CAT Application

Each examinee was given a CAT of a fixed length, with items selected from among the 300 operational items. After the adaptive part of the test, all respondents were given the same 23 pretest items. For implementing the adaptive part, the real trait levels of the examinees were drawn from a standard normal distribution. Maximum a posteriori (MAP) was used for trait-level estimation, with a standard normal as prior distribution. Items were selected using the progressive method (Revuelta & Ponsoda, 1998). In this method, the selection of items has a high random component at the start of the test, while the importance of Fisher's information increases as the test advances. It has been suggested that this item selection method should be among those preferred when item bank security is a concern (Barrada, Olea, Ponsoda, & Abad, 2010).

It must be noted that the pretest items were tested for DIF one at a time, excluding the information of the remaining pretest items. In these conditions, using a large number of pretest items is equivalent to using a single pretest item in a within-subjects design. In this way, the simulation time and sample variability are reduced.

Parameters of the Operational Items

The Law School Admission Test (LSAT) parameter distributions were used to create the 300 operational items, following the information provided by Nandakumar and Roussos (2004). For items with a *b* parameter ≤ –1, the *a* parameter followed a log-normal distribution (–0.357, 0.25) within the [0.4, 1.1] range. For the rest of the items, the *a* parameter distribution was log-normal (–0.223, 0.34) within the [0.4, 1.7] range. The *b* parameter followed a *N*(0, 1) distribution within the [–3, 3] range, and the *c* parameter followed a uniform distribution within the [.12, .22] range. The operational item parameters were the same for the reference and focal groups (i.e., no DIF was considered for these items).

Parameters of the Pretest Items

The number of pretest items was 23. Ten of these had unidirectional DIF (i.e., the item systematically favors one group), six had non-unidirectional DIF (i.e., the item favors one group or the other depending on the ability level), and seven had no DIF. Two conditions in DIF size were simulated: Half of the items had moderate DIF (β = .05), whereas the other half had large DIF (β = .10). DIF size was classified as moderate or large following Dorans and Holland's (1993) criteria. β is a measure of the magnitude of the effect of DIF (Wainer, 1993), calculated as

$$\beta = \int |P(u_{iR} = 1 \mid \theta) - P(u_{iF} = 1 \mid \theta)| g(\theta), \qquad (5)$$

where $P(u_{iR} = 1 \mid \theta)$ and $P(u_{iF} = 1 \mid \theta)$ denote the probability of answering the item correctly for the reference and focal groups, respectively, and $g(\theta)$ is the density of θ, assuming a standard normal distribution. In other words, β is the expected value across ability levels of the absolute difference in the probability of correct responses between the groups when both groups have a standard normal ability distribution.

- Items with unidirectional DIF. First, the items were produced for the focal group. Subsequently, the *b* parameter in the reference group was adjusted to obtain the desired DIF size. Of the 10 items, eight were obtained from the combination of the value in the *a* parameter (0.7 or

1.2), the value in the *b* parameter (–1.3 or 1.3), and the size of the DIF. The other two items represented an average LSAT item ($a = 0.8$, $b = 0$) with two different DIF sizes. The items were easier for the reference group because their *b* parameter was lower.

- Items with non-unidirectional DIF. The *a* parameter for the focal group was equal to 0.8 for all these items. Non-unidirectional DIF was created by changing the *a* parameter in the reference group to obtain the desired DIF size. The six items were obtained from the combination of the value in the *b* parameter (–1.3, 0, 1.3) and the size of the DIF. The items were more discriminative for the reference group because their *a* parameter was higher.
- Items without DIF. Seven items were used, of which six were obtained from the combination of the value in the *a* parameter (0.7, 0.8 or 1.2) and the value in the *b* parameter (–1.3 or 1.3). The other item represented an average LSAT item (a = 0.8; b = 0).

For all the pretest items, the *c* parameter was set to .17. The parameter values of the items studied were considered to be within the range observed in the actual tests (Lopez-Rivas, Stark, & Chernyshenko, 2008). Table 1 shows the item parameters.

Table 1. Item parameters used to generate the studied items.

No-DIF			Unidirectional DIF					Non-Unidirectional DIF				
Item	$a_R=a_F$	$b_R=b_F$	Item	$a_F=a_R$	b_F	b_R	β	Item	a_F	a_R	$b_R=b_F$	β
1	0.70	–1.30	8	0.70	–1.30	–1.69	.05	18	0.80	1.23	–1.30	.05
2	0.70	1.30	9	0.70	1.30	0.97	.05	19	0.80	1.23	1.30	.05
3	1.20	–1.30	10	1.20	–1.30	–1.685	.05	20	0.80	1.19	0.00	.05
4	1.20	1.30	11	1.20	1.30	1.005	.05	21	0.80	2.42	–1.30	.10
5	0.80	–1.30	12	0.80	0.00	–0.24	.05	22	0.80	2.41	1.30	.10
6	0.80	1.30	13	0.70	–1.30	–2.2	.10	23	0.80	1.83	0.00	.10
7	0.80	0.00	14	0.70	1.30	0.675	.10					
			15	1.20	–1.30	–2.325	.10					
			16	1.20	1.30	0.755	.10					
			17	0.80	0.00	–0.49	.10					

Note. The *c* parameter = .17 for all the items.

Manipulated Factors

The independent variables were as follows:

- CAT length: Two test lengths of 10 and 30 items, respectively, were simulated. CAT items were selected from the operational item pool.
- Impact: Whereas the trait level for the reference group always followed $N(0, 1)$, three conditions were generated for the focal group: without impact [$N(0, 1)$], with unfavorable impact [$N(–0.5, 1)$], and with favorable impact [$N(0.5, 1)$]. These distributions are common in applied contexts and simulation studies (French & Finch, 2008; Finch & French, 2007; Lei et al., 2006; Stark, Chernyshenko, & Drasgow, 2006).
- Sample size: Two equal sample-size conditions were included for the focal/reference groups in the study: 250/250 and 500/500. A sample size of 500 examinees per group is recommended for the DIF analysis in fixed tests in procedures based on the IRT (Clauser & Mazor, 1998). The small samples were close to those used by Kim (2006) as an example of a realistic situation within the context of online calibration.

Each of the 12 different conditions (2×3×2) was replicated 100 times. In each repetition, the 23 pretest items were administered varied according to the following factors:

- Type of DIF: Without DIF, unidirectional DIF, and non-unidirectional DIF.
- DIF size: Small ($\beta = .05$) and large ($\beta = .10$).

An additional manipulated factor was the use of empirical or imputed data sets (see the following paragraph).

Compared Methods for DIF Detection

Two methods were compared: (a) IRT-LRT with FIPC and imputation, and (b) IRT-LRT with FIPC and without imputation. The methods were programmed using the subroutines provided by the computer program ICL (Hanson, 2002). To test the DIF, the pretest items were evaluated independently one by one, using just the operational items as anchors and excluding the responses to the remaining pretest items. To obtain each IRT-LRT contrast (Equation 1), FIPC was applied twice to get the log-likelihood: First, for the compact model (assuming that the pretest item parameters are equal across the groups) and, second, for the augmented model (allowing the studied item parameters to differ). The difference in the number of estimated parameters was three (the number of item parameters), which was the degrees of freedom of the IRT-LR statistics.

In applying FIPC, the MWU-MEM was used (see Kim, 2006). In this method, the operational and pretest item parameters were used to compute and update the posterior probabilities in the *E* step (Equation 3). In both models (augmented and compact), the operational item parameters were fixed to their known values, and the pretest item parameters were estimated. The prior distributions for the *a* and *c* pretest item parameters were defined {$a \sim$ log-normal (0, 0.5); $c \sim$ beta (4, 16)}. These distributions are common to other estimation programs, such as PARSCALE and BILOG (du Toit, 2003).

In the augmented model, the ability distribution parameters were freely estimated in the reference and focal groups. The obtained ability distribution parameters were used (and fixed) in the compact model to preserve nesting for the IRT-LRT (a similar approach was used by Woods, 2008). The ICL syntax and programs for automating the process are available from the authors.

In the method without imputation, FIPC analysis was applied to the incomplete data set (with missing values for the non-administered operational items). In the method with imputation, missing responses were previously imputed before applying FIPC analysis. First, MAP ability estimates were computed based only on responses to operational items applied in the CAT. Then, probabilities of getting the items correct were generated according the known operational item parameters and CAT ability estimates. Finally, the imputed response was set equal to 1 if the probability value was greater than a random number from a uniform (0, 1) distribution and equal to 0 otherwise.

Evaluation Criteria

The relative effectiveness of the two methods (with and without imputation) was evaluated according to the following criteria:

- DIF detection: To evaluate the effectiveness of DIF detection, the Type I error and power were calculated. A critical value was established for Type I errors of .05. The power was regarded as acceptable when its value was $\geq .80$.
- Recovery of distribution parameters for each group: The estimated mean and standard deviation of θ for the reference and focal groups were obtained in each augmented model. These values were compared with the true known values. The degree of recovery for these parameters was evaluated through bias.

- Recovery of DIF size (β): The quality of β recovery was evaluated through bias. Following Nandakumar and Roussos (2004), we considered a bias > .01 in the effect size as problematic. $\hat{\beta}$ was estimated from Equation 5, using the pretest item parameters estimated in the augmented model; $g(\theta)$ denoted the standard normal distribution. The estimated ability distributions were not considered in computing $\hat{\beta}$. Thus, differences between β and $\hat{\beta}$ should be attributed to differences between the true and the estimated item pretest parameters.

Analysis

ANOVAs were performed to determine the factors that affected our evaluation criteria. A mixed ANOVA was used, in which the methods (with and without imputation) applied were regarded as within-simulees variables, whereas CAT length, sample size, impact, and DIF effect size were treated as between-simulees variables. These analyses were carried out separately for unidirectional and non-unidirectional DIF.

To gauge the effect size, a partial η^2 measure was used. Reported results correspond only to effects with a magnitude of more than .06. Partial η^2 over this value are considered as medium and over .14 as large in J. Cohen's classification (1988, 1992). All the reported effects were significant at $p < .05$.

RESULTS

Type I Error Rate

Table 2 presents the proportion of false positives (Type I error rates) for each type of method (non-imputed vs. imputed) by sample size, test length, and impact. It can be observed that the error rates are below the nominal level (i.e., < .05) across all the conditions. An ANOVA showed that the Type I error rates were essentially unaffected by the independent variables. None of these variables or interactions demonstrated a moderate or higher effect size ($\eta^2 > .06$).

Power

Power for unidirectional DIF. Table 3 (top) shows the power in detecting unidirectional DIF. As expected, the power increases with the DIF size [$F(1, 2376) = 5615.27$, $\eta^2 = .703$] and sample size [$F(1, 2376) =$

987.83, η^2 = .294]. Surprisingly, increments in the test length had no significant effect on power. Adequate power was achieved when the sizes of the DIF and sample were large. With regard to the methods, and contrary to our hypothesis, the non-imputation alternative was not found to be superior to the imputation method. An interaction was found between the methods and the θ distribution of the focal group [$F(2, 2376) = 478.67$, η^2 = .287]. Also, there was an almost medium sized interaction effect among the methods, θ distribution, and test length [$F(2\ .2376) = 72.95$, η^2 = .058], as shown in Figure 1. As can be seen from the figure, non-imputation conditions were almost unaffected by impact. However, when missing values were imputed, the power was reduced because the impact shifted from favoring the reference group to the reverse. When the impact favored the reference group [$N_F(-0.5, 1)$], imputation performed better than the non-imputation method. In the absence of impact, there was no difference between the methods. Finally, when the impact favored the focal group [$N_F(0.5, 1)$], the imputation method performed worse than the non-imputation method. These effects were higher when the test was short.

Table 2. Type I error rate by sample size, test length, θ distribution of the focal group and method.

Sample sizes	Test length	$N(-0.5,1)$		$N(0,1)$		$N(0.5,1)$	
		Non-imputed	Imputed	Non-imputed	Imputed	Non-imputed	Imputed
250/250	10	.019	.030	.013	.014	.030	.039
	30	.021	.029	.031	.031	.027	.031
500/500	10	.019	.043	.029	.027	.020	.044
	30	.020	.023	.027	.030	.030	.031

Power for non-unidirectional DIF. Table 3 (bottom) shows the power for the non-unidirectional DIF. It can be noted that the pattern of results is similar to that described for unidirectional DIF; i.e., the power increases with the DIF size [$F(1, 2376) = 3528.10$, η^2 = .598] and sample size [$F(1, 2376) = 744.70$, η^2 = .239]. Adequate power was achieved when

the sizes of the DIF and sample were large. For non-unidirectional DIF, the presence of imputed responses had no relevant main or interaction effects (see Figure 1). Again, non-imputed conditions did not perform better than conditions with imputations.

Table 3. Power rate by sample size, test length, θ distribution of the focal group and method for both unidirectional and non-unidirectional DIF.

β	Sample size	Test length	N(-0.5,1)		N(0,1)		N(0.5,1)	
			Non-imputed	Imputed	Non-imputed	Imputed	Non-imputed	Imputed
					Unidirectional			
.05								
	250/250	10	.166	.278	.146	.154	.150	.086
		30	.166	.202	.170	.166	.164	.126
	500/500	10	.340	.536	.360	.344	.322	.162
		30	.322	.422	.358	.366	.338	.266
.10								
	250/250	10	.632	.756	.690	.688	.634	.510
		30	.636	.678	.656	.656	.644	.584
	500/500	10	**.866**	**.948**	**.906**	**.910**	**.914**	**.834**
		30	**.890**	**.928**	**.928**	**.924**	**.912**	**.884**
					Non-Unidirectional			
.05								
	250/250	10	.153	.163	.137	.143	.203	.137
		30	.173	.177	.190	.190	.207	.180
	500/500	10	.310	.353	.370	.380	.393	.327
		30	.387	.370	.413	.403	.497	.473
.10								
	250/250	10	.557	.573	.623	.633	.703	.713
		30	.663	.670	.707	.707	.717	.730
	500/500	10	**.870**	**.837**	**.897**	**.903**	**.897**	**.923**
		30	**.913**	**.897**	**.973**	**.957**	**.977**	**.973**

Note. Bold cells correspond to power rates equal or greater than .80.

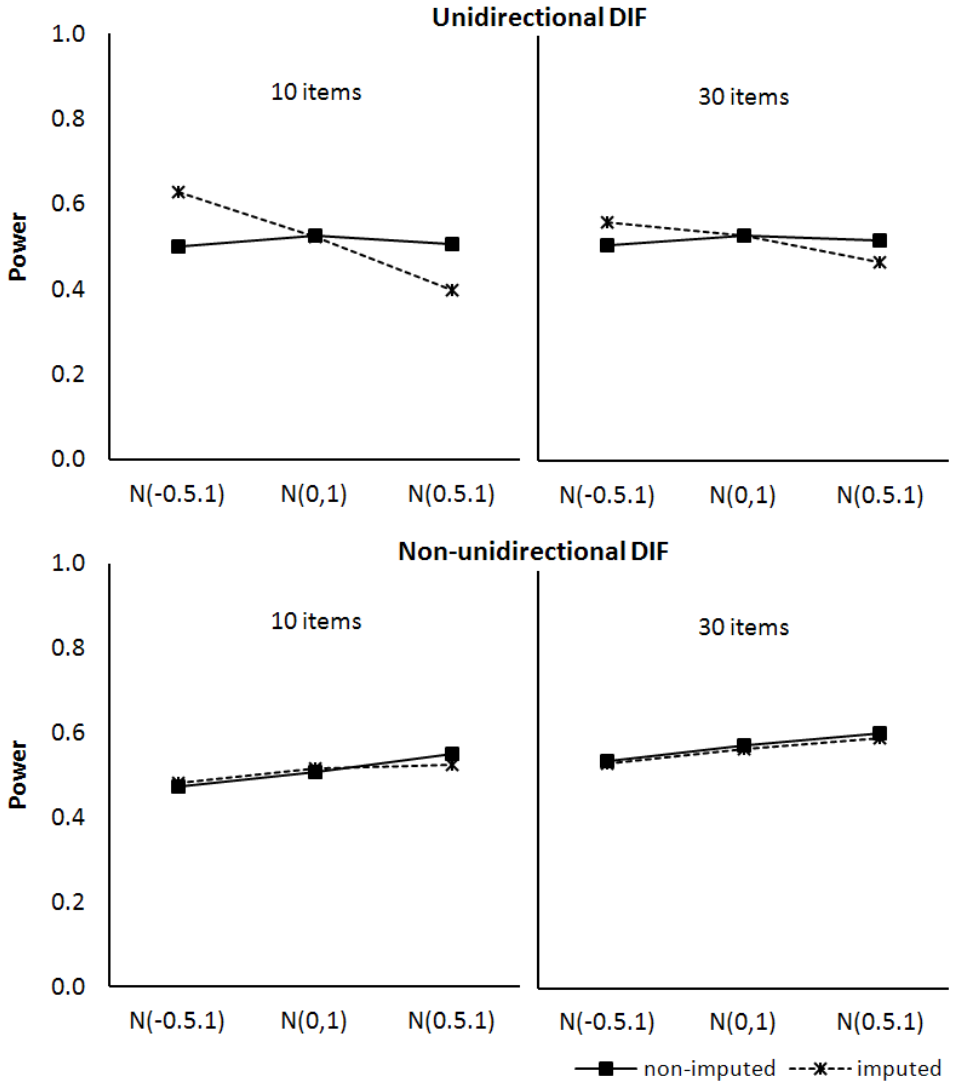

Figure 1. Power rates for detecting unidirectional and non-unidirectional DIF by focal group θ distribution, test length and method.

Recovery of the ability distribution parameters. Figures 2 and 3 present a summary of the results of the bias in the ability distribution parameters recovery. It can be observed that there is no appreciable difference in the recovery of the distribution by type of DIF, DIF size, or

sample size. Therefore, the average bias for the mean (Figure 2) and standard deviation (Figure 3) for each method has been plotted as a function of the test length and focal group θ distribution.

The average bias for the reference group mean (Figure 2, top) was higher when responses were imputed [$F(1, 1188) = 7456.12$, $\eta^2 = .863$]. Imputation overestimated the mean, whereas non-imputed conditions led to unbiased estimates. An interaction effect was found between the method and test length [$F(1, 1188) = 1627.98$, $\eta^2 = .578$] because the overestimation was greater in the short-test condition.

The bias of the focal group mean (Figure 2, bottom) varied according to the presence of imputed responses [$F(1, 1188) = 7904.16$, $\eta^2 = .869$], with imputation producing a higher bias. Non-imputation led to unbiased results. The bias for the imputation method varied with the impact, as indicated by a two-way interaction between the method and the focal group distribution [$F(2, 1188) = 13755.05$, $\eta^2 = .959$], and with the test length [$F(1, 1188) = 1902.01$, $\eta^2 = .616$]. Finally, a three-way interaction was found between the method, distribution, and test length [$F(1, 1188) = 3486.57$, $\eta^2 = .854$]. The imputation method produced a positive bias (an overestimation) of the underlying mean at $N_F(–0.5, 1)$, which decreased when $N_F(0, 1)$ and became negative (underestimate) at $N_F(–0.5, 1)$; this effect was higher with a short test.

The analysis of bias of the standard deviation revealed a large size effect for method [reference group: $F(1, 1188) = 25276.56$, $\eta^2 = .955$; focal group: $F(1, 1188) = 25498.07$, $\eta^2 = .955$] and test length [reference group: $F(1, 1188) = 329.42$, $\eta^2 = .217$; focal group: $F(1, 1188) = 282.93$, $\eta^2 = .192$] for both groups. There was also an interaction between method and test length for both groups [reference group: $F(1, 1188) = 9039.52$, $\eta^2 = .884$; focal group: $F(1, 1188) = 8654.48$, $\eta^2 = .879$]. An important underestimation of the standard deviation was detected when using imputation when compared with non-imputation, and this tended to be larger when the test length was shorter. For the focal group, there was also a triple interaction between method, test length, and impact [$F(2, 1188) = 36.97$, $\eta^2 = .059$]. The graph presented in Figure 3 shows a larger underestimation for imputation when a test length of 10 items and the distribution $N_F(–0.5, 1)$ were used. In brief, non-imputation had better scale recovery for both groups in all the conditions.

Recovery of the DIF effect size. An additional issue pertinent to understanding the performance of the methods is effect size recovery. Figure 4 shows the average bias for unidirectional (top) and non-

unidirectional (bottom) for β in both methods as a function of the focal group θ distribution and test length. The bias for the non-imputed conditions remained stable and, importantly, below the limit of .01 (Nandakumar & Roussos, 2004) for different lengths and impacts. For the unidirectional DIF, a large three-way interaction was found between method, distribution, and test length [$F(2, 2376) = 1500.439$, $\eta^2 = .558$]. The bias for the imputed method ranged from positive and above the bias of the non-imputed methods when $N_F(-0.5, 1)$ to negative when $N_F(0.5,1)$. This trend was more pronounced when the test was short. A second three-way interaction was found between method, distribution, and DIF size [$F(1, 2376) = 87.55$, $\eta^2 = .069$]. In this case, the interaction between method and distribution was similar to that previously described but modulated by the DIF size, as shown in Figure 5 (top).

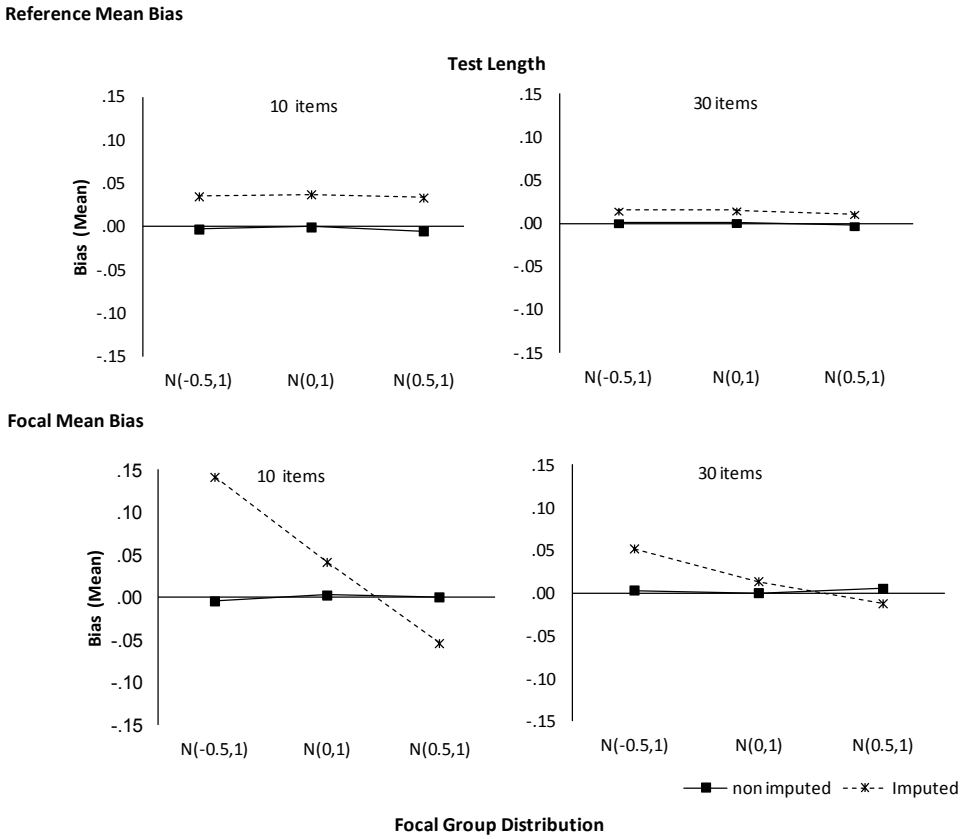

Figure 2. Average bias of the mean for the reference (top) and the focal (bottom) groups by test length and focal group θ distribution.

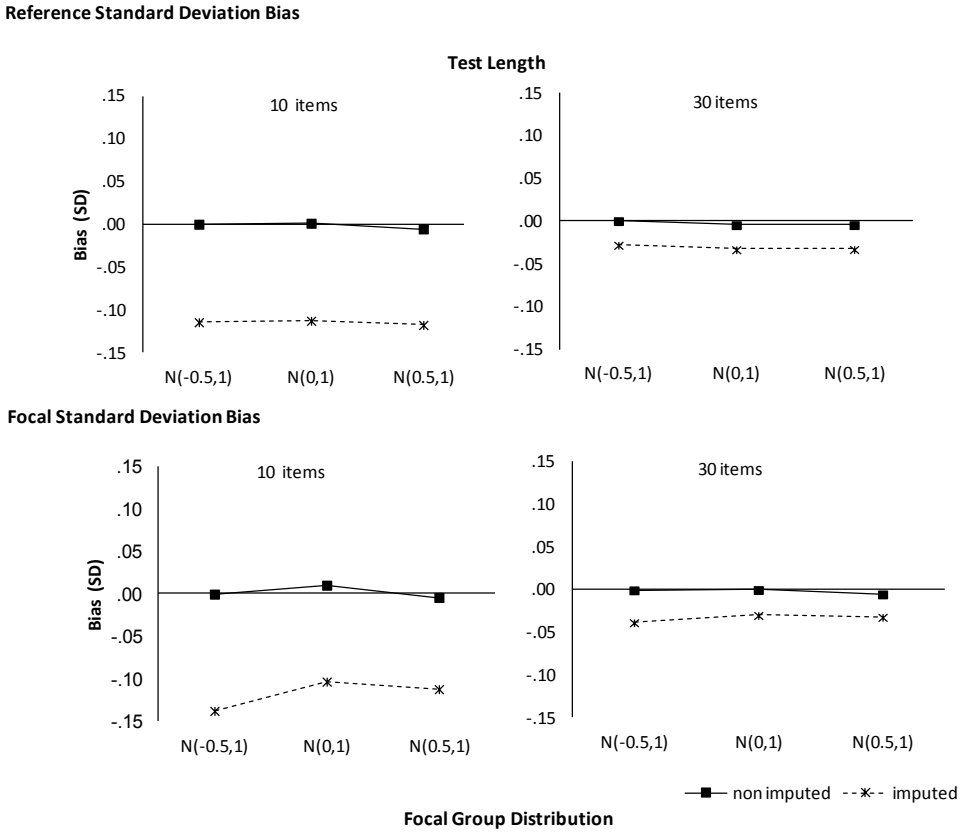

Figure 3. Average bias of the standard deviation for the reference (top) and the focal (bottom) groups by test length and focal group θ distribution.

For the non-unidirectional DIF (Figure 4, bottom), imputation produced an underestimation of β [$F(1, 2376) = 3767.33$, $\eta^2 = .613$]. There was an increment in the effect size underestimation bias when the DIF size was large (see Figure 5, bottom) [$F(1, 2376) = 627.75$, $\eta^2 = .209$]. Finally, an interaction effect of method with test length [$F(1, 2376) = 409.87$, $\eta^2 = .147$] and DIF effect size [$F(1, 2376) = 607.36$, $\eta^2 = .204$] was found, implying that the differences between the methods were higher in large DIF conditions and with short tests (see Figure 6). The average underestimation was > .01 for imputation with large DIF and a test length of 10 items.

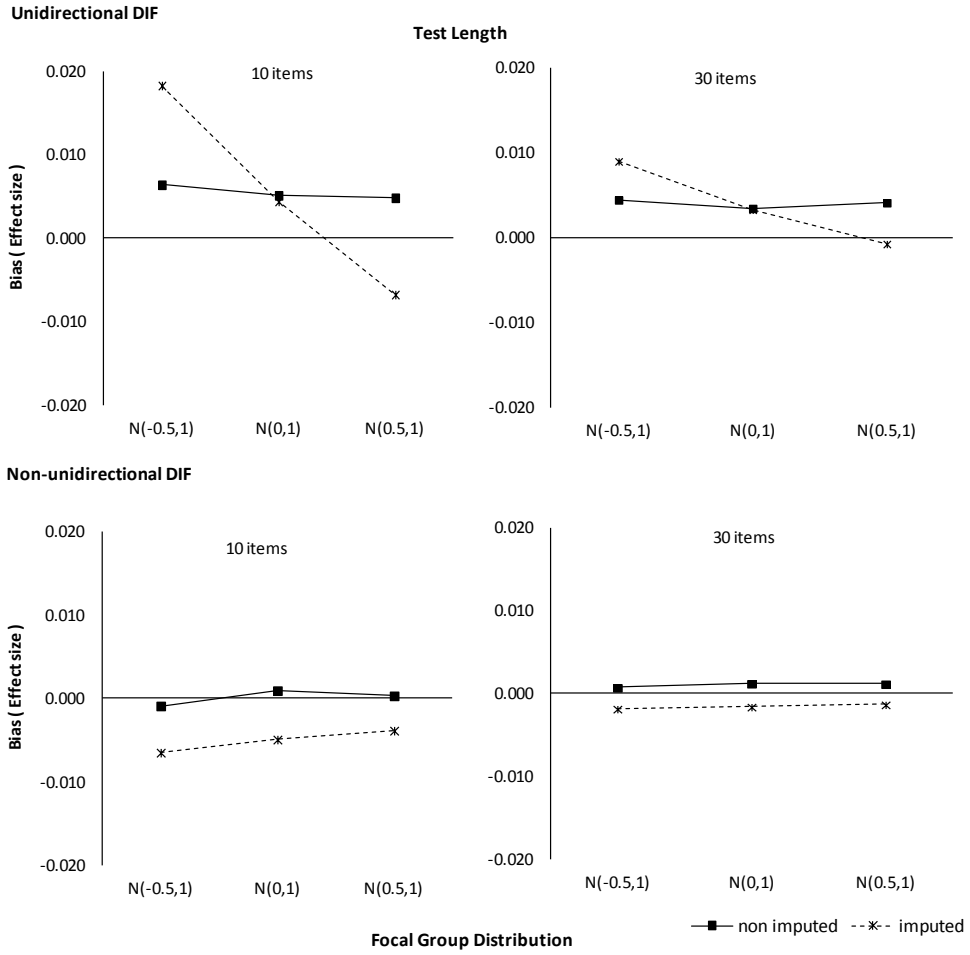

Figure 4. Average bias of the effect size for unidirectional DIF (top) and non-unidirectional (bottom) for each method by test length and focal group θ distribution.

DISCUSSION AND CONCLUSIONS

In the present study, we evaluated the feasibility of applying FIPC in the IRT-LRT for DIF analysis for pretest items. We consider FIPC as conceptually more efficient for the IRT-LRT than the concurrent calibration used by Lei et al. (2006) due to several reasons. First, FIPC is more time-efficient and avoids problems related to incompleteness of the response matrix because it does not require reestimation of operational items.

Second, multigroup MWU-MEM FIPC can be easily applied using the subroutines provided by the computer program ICL (Hanson, 2002), which, unlike IRTLRDIF, does not have restrictions in the size of the matrix to be analyzed (Lei et al., 2006). Third, the estimated parameters of the new items are automatically placed on the existing scale; therefore, a linking step is not required.

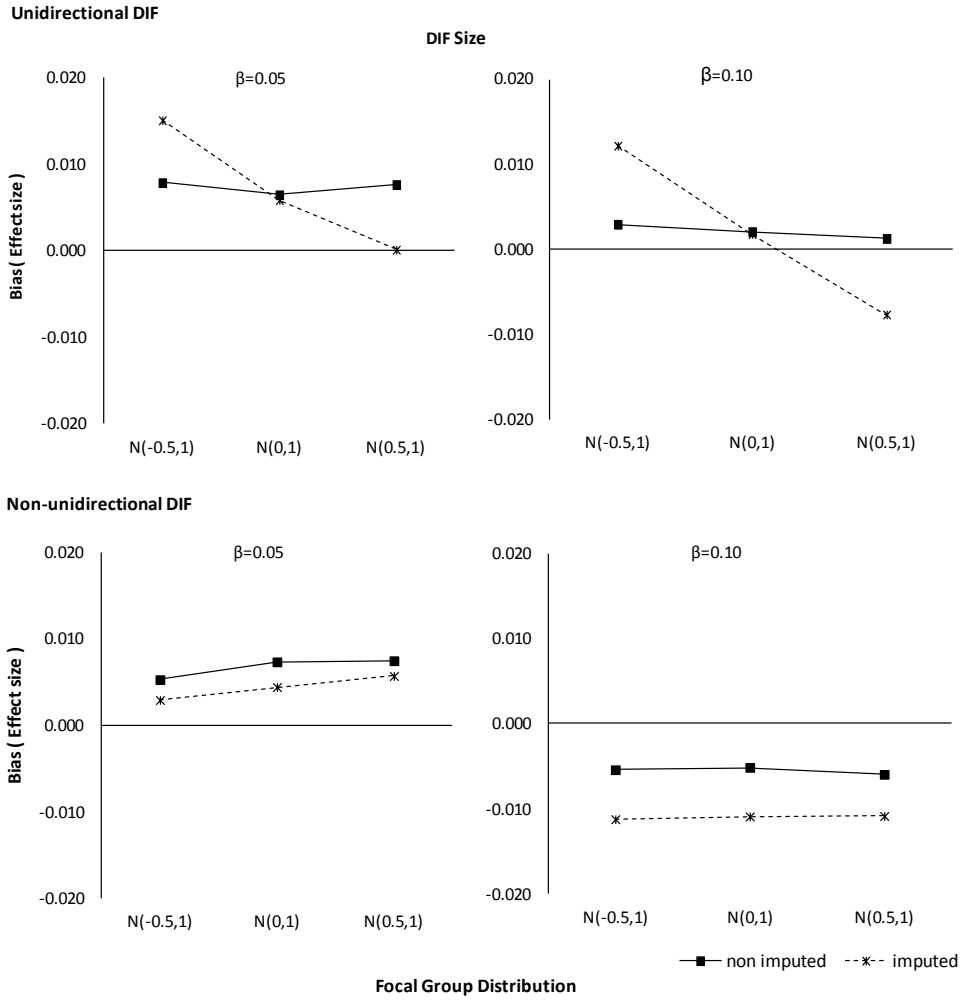

Figure 5. Average bias of the effect size for unidirectional DIF (top) and non-unidirectional (bottom) for each method by DIF size and focal group θ distribution (unidirectional DIF).

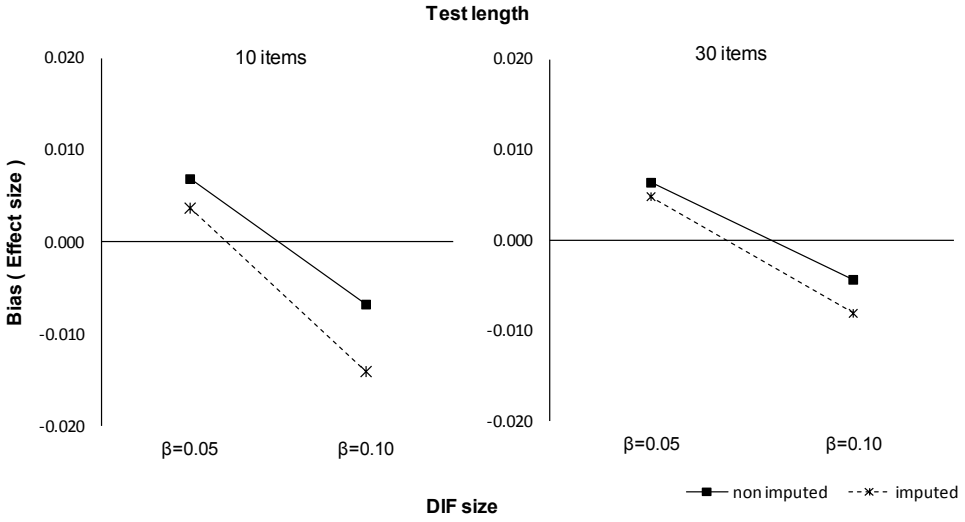

Figure 6. Average bias of the effect size for each method by test length and DIF size (non-unidirectional DIF).

Lei et al. (2006) proposed the imputation of missing responses by taking the estimated trait level as a real trait level. To test the effects of response imputation, we applied FIPC with and without imputation. Type I error rates, power, recovery of latent ability distributions, and recovery of the effect size were evaluated across several simulated conditions. The results suggest that with both methods the Type I error rates are clearly under the nominal level. Overall Type I error rates were .027, while α was equal to .05. This indicates that the test statistic does not follow its theoretical distribution. These findings agree with those obtained in conventional tests (Cohen, Kim, & Wollack, 1996; Finch, 2005; Finch & French, 2007; Kim & Cohen, 1995, 1998; Lopez-Rivas et al., 2008). In fact, Lopez-Rivas et al. (2008) reported Type I error rates of < .01 when the anchor test was optimal. It could be argued that in applied settings it could be possible to use nominal values over .05 to get actual errors near that desired level. As none of the manipulated variables showed a moderate or higher effect on Type I error rates, it seems that the α value that leads to Type I error rates equal to .05 could be fixed with independence of the conditions of the CAT.

On the other hand, it was found that regardless of the method, power rates were impacted greatly by the DIF size and sample size, as reported in previous DIF studies with conventional (Finch, 2005; Finch & French, 2007; Stark et al., 2006; W. Wang, 2004; W. Wang & Yeh, 2003) and adaptive tests (Lei et al., 2006; Nandakumar & Roussos, 2001, 2004). In the present study, samples of 1,000 examinees were required to achieve suitable power values (> .80). In addition, these values were only reached in the detection of items with large DIF.

Despite the above-mentioned similarities, some differences between the methods were found in the power for detecting unidirectional DIF in the presence of impact. In this condition, when responses were imputed, the power rates were severely dependent on the ability distributions of the focal group, especially when the CAT length was shorter. Imputation showed a lower power rate compared to non-imputation in the $N(-0.5, 1)$ condition but, unexpectedly, was superior in the $N(0.5,1)$ condition.

This unexpected last result is actually due to an overestimation of the effect size ($\hat{\beta}$) produced by a severe bias in the scale recovery. An in-depth analysis revealed that imputation produces compression in the scale (i.e., the standard deviation of the latent trait distribution had been underestimated for both groups) and that, in the presence of impact, the latent trait mean of the focal group moves toward zero (it yields to an underestimation when $\mu_F = -0.5$ and to an overestimation when $\mu_F = 0.5$). These distortions are greater when the CAT is short, and they occur due to the fact that the estimated MAP ability level is used to carry out the imputations. It is well known that the MAP procedure produces biased estimates toward the mean of the prior distribution and that the bias tends to correct itself as the length of the test increases (T. Wang & Vispoel, 1998). Thus, because imputed responses are generated with the CAT ability MAP estimate, which are shrunk toward the mean of the prior distribution (zero), the latent ability distribution of the focal group is also shrunk. Thus, in the $N(-0.5, 1)$ condition, as an effect of imputation, the mean for the focal group is overestimated. Consequently, the difficulty parameter of the item in the focal group is also overestimated. It must be noted that in our simulation study, in the unidirectional DIF conditions, the true b_F parameters were higher than the true b_R parameters (the items were easier for the reference group). Overestimating the b_F parameters artificially increases the difference between the item characteristic curves of the groups, causing the effect size to be overestimated and thus increasing the DIF detection rate.

Indeed, the results indicate that FIPC without imputation shows more precise distribution parameter and $\hat{\beta}$ estimates, regardless of the factors manipulated. The quality of the recovery does not depend on the impact or test length. Moreover, the use of the pretest item with DIF for updating posterior probabilities does not seem to have affected FIPC negatively. In all the simulated conditions without imputation, the average bias in estimating the effect size was never > .01, whereas in some conditions with imputed responses, this limit was surpassed.

If we take into account the results regarding the factors and efficiency of the methods, we can conclude that multigroup FIPC is a feasible procedure for assessing DIF in pretest items applied in online calibration settings. FIPC shows power rates over .557 when the DIF size was large (β = .10), even with small sample sizes (N = 250/250). At this point, the known trade-off between Type I error rates and power should be noted. Higher power could be achieved if Type I errors approached their nominal level. Power results are encouraging because they indicate that the technique can be used even in the first phases of studies of the psychometric properties of pretest items.

We can conclude that it is not recommendable to impute the responses in the data matrix generated by a CAT, especially for short CAT. Single imputation based on Bayesian CAT ability estimate, results in biased standard deviations for the latent trait distributions. Moreover, in the presence of impact, biased impact and DIF size effect measures may be obtained. This is important because the decision to discard an item due to the presence of DIF should be based both on the measurement of statistical significance and on the evaluation of the magnitude of the DIF. Furthermore, the recovery of latent distribution parameters is important not only to study impact in the CAT environment but also to obtain DIF size effect measures based on the focal group distribution (Wainer, 1993).

Some issues deserve attention in subsequent studies. Following the results of previous simulation studies (Ban et al., 2001, 2002; Kim, 2006), the MWU-MEM was judged as the best way to conduct FIPC, although in those studies no item had DIF. Future studies should consider the effect of DIF in the different available FIPC methods. In our study, only one pretest at a time was analyzed, excluding the remaining pretest items. We preferred this approach because the inclusion of other (with suspicious DIF) pretest items could affect parameter estimates with the MWU-MEM. Future research should compare the performance of the methods when more than one pretest item is included. Additional studies should extend the FIPC to

the detection of DIF in operational items. Finally, the discrepancy between expected and obtained Type I errors deserves further research.

Although FIPC provides one easy method for testing DIF, other procedures for imputing responses deserve more attention. Recently, Finch (2011) has suggested the use of Multiple imputation for detecting DIF in fixed tests, when researcher is interested in item parameters. Additionally, correcting Bayesian estimates for unreliability might be considered (de la Torre & Deng, 2008).

Summarizing, we have shown how FIPC could be implemented to detect DIF using IRT-LRT in CAT environments. This has been done with ICL (Hanson, 2002), a freeware program. Contrary to previous claims (Lei et al., 2006), the sparse matrix, characteristic of CATs, does not need to be filled with imputed responses generated with the CAT ability estimate. In fact, non-imputing responses, when compared with imputation, is a better option in terms of power (equal mean power, but this power does not depend on specific simulation conditions) and recovery of parameters of the trait level distribution and effect size.

RESUMEN

Calibración con parámetros de los ítems fijos para la evaluación del funcionamiento diferencial del ítem en tests adaptativos informatizados. En tests adaptativos informatizados los ítems pretest se presentan junto con los ítems operativos para renovar el banco de ítems. Los ítems pretest se calibran y se analiza el posible funcionamiento diferencial de los ítems (FDI). Este análisis presenta algunos problemas debido a la gran cantidad de respuestas faltantes, una de las posibles soluciones es el uso de métodos de calibración con parámetros fijos (Kim, 2006). En este estudio, aplicamos el método de múltiples actualizaciones de los pesos y múltiples ciclos EM con imputación de respuestas (tal y como propusieron Lei, Chen, y Yu, 2006) y sin imputación de respuesta para los ítems no aplicados. Empleamos el test de razón de verosimilitudes de la TRI para la detección del FDI. Los factores manipulados fueron el tipo de FDI, el tamaño del FDI, el tamaño del impacto, la longitud del test, y el tamaño de las muestras. Los resultados señalan que el método de calibración con parámetros fijos es una alternativa adecuada para la detección de un FDI grande cuando se utilizaron muestras grandes. En presencia de impacto el uso de imputación de respuestas introdujo un sesgo en las medidas del tamaño del efecto del FDI.

REFERENCES

Ban, J. C., Hanson, B. A., Wang, T., Yi, Q., & Harris, D. J. (2001). A comparative study of on-line pretest-item calibration/scaling methods in computerized adaptive testing. *Journal of Educational Measurement, 38*, 191-212.

Ban, J. C., Hanson, B. A., Yi, Q., & Harris, D. J. (2002). Data sparseness and on-line pretest item calibration-scaling methods in CAT. *Journal of Educational Measurement, 39*, 207-218.

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. (2010). A method for the comparison of item selection rules in computerized adaptive testing. *Applied Psychological Measurement, 34*, 438-452.

Cohen, A., Kim. S., & Wollack, J. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement, 20,* 15-26.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). Hillsdale, NJ. Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155-159.

Clauser, B., & Mazor, K., (1998) Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17,* 31-44.

de la Torre, J., & Deng, W. (2008). Improving person fit assessment by correcting the ability estimate and its reference distribution. *Journal of Educational Measurement, 45*, 159-177.

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

Dorans, N., y Holland, P. (1993). *DIF detection and description: Mantel–Haenszel and standardization*. In P. W. Holland y H. Wainer (Ed.), Differential item functioning (pp.35–66). Hillsdale, NJ: Erlbaum.

du Toit, M. (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Lincolnwood, IL: Scientific Software International.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29*, 278-295.

Finch, W. H., & French, B. (2007). Detection of crossing differential item functioning: A comparison of four methods. *Educational and Psychological Measurement, 67*, 565-582.

French, B., & Finch, H. (2008). Multigroup confirmatory factor analysis: Locating the invariant referent sets. *Structural Equation Modeling, 15*, 96-113.

Finch, H. (2011). The use of multiple imputation for missing data in uniform DIF analysis: Power and Type I error rates. *Applied Measurement in Education*, *24*, 281-301.

Hanson, B. A. (2002). *IRT Command Language (Version 0.020301)*. Monterey, CA: Author. (Available at http://sourceforge.net/projects/ssm)

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement*, *43*, 355-381.

Kim, S., & Cohen, A. S. (1995). A comparison of Lord's chi-square, Raju's area measures, and the likelihood ratio test on detection of differential item functioning. *Applied Measurement in Education, 8,* 291-312.

Kim, S., & Cohen, A. S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement,*

22, 345-355.Haynie, K. A., & Way, W. D. (1995). *An investigation of item calibration procedures for a computerized licensure examination*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Harmes, J. C., Parshall, C. G., & Kromrey, J. D. (2001). *Online item parameter recalibration: Application of missing data treatments to overcome the effects of sparse data conditions in a computerized adaptive version of the MCAT*. Report submitted to the Association of American Medical Colleges, Section for the MCAT.

Harmes, J. C., Parshall, C. G., & Kromrey, J. D. (2003). *Recalibration of IRT item parameters in a CAT: Sparse data matrices and missing data treatments*. Paper presented at the annual meeting of the National Council on Measurement in Education. Chicago: IL.

Hsu, Y., Thompson, T. D., & Chen, W.-H. (1998). *CAT item calibration*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego.

Lei, P. W., Chen, S. Y., & Yu, L. (2006). Comparing methods of assessing differential item functioning in a computerized adaptive testing environment. *Journal of Educational Measurement, 43*, 245–264.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley & Sons.

Lopez-Rivas, G., Stark, S., & Chernyshenko, O. (2008). The effects of referent item parameters on differential item functioning detection using the free baseline likelihood ratio test. *Applied Psychological Measurement, 33*, 251-265.

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

Mills, C. N., & Stocking, M. L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education, 9*, 287-304.

Miller, T. R. (1992). *Practical considerations for conducting studies of differential item functioning (DIF) in a CAT environment*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Nandakumar, R., & Roussos, L. (2001). *CATSIB: A modified SIBTEST procedure to detect differential item functioning in computerized adaptive tests* (Research report). Newtown, PA: Law School Admission Council.

Nandakumar, R., & Roussos, L. A. (2004). Evaluation of the CATSIB DIF procedure in a pretest setting. *Journal of Educational and Behavioral Statistics, 29*, 177–200.

Parshall, C. G. (1998). *Item development and pretesting in a computer-based testing environment*. Paper presented at the colloquium Computer-Based Testing: Building the Foundation for Future Assessments, Philadelphia, PA.

Pommerich, M. & Segall, D.O. (2004). Calibrating CAT pools and online pretest items using marginal maximum likelihood methods. Presented at Annual Meeting of the National Council on Measurement in Education. Retrieved March 4, 2014 from http://www.editlib.org/p/97315.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, *35*, 311-327.

Steinberg, L., Thissen, D. & Wainer, H. (1990). Validity. En Wainer (Ed.), *Computer adaptive testing: A primer*. (pp. 185-230). Hillsdale, NJ: Lawrence Erlbaum

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, *91*, 1291-1306.

Stocking, M. L. (1988). *Scale drift in on-line calibration* (ETS Research Report 88-28). Princeton, NJ: ETS.

Thissen, D. (2001). *IRTLRDIF v2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning* [Computer software and manual]. Chapel Hill: L.L. Thurstone Psychometric Laboratory, University of North Carolina.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds), *Test validity (*pp. 147-69). Hillsdale, NJ: Lawrence Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of the estimated IRT models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Erlbaum.

Wainer, H. (1993). Model-based standardized measurement of an item differential impact. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 123-135). Hillsdale, NJ: Erlbaum.

Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement, 35*, 109-135.

Wang, W. (2004). Effects of anchor item methods on detection of differential item functioning within the family of Rasch models. *Journal of Experimental Education, 72*, 221-261.

Wang, W., & Yeh, Y. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement, 27*, 479-498.

Woodruff, D. J., & Hanson, B. A. (1996). *Estimation of ítem response models using the EM algorithm for finite mixtures* (ACT Research Report 96-6). Iowa City, IA: ACT, Inc.

Woods, C. M. (2008). IRT-LR-DIF with estimation of the focal-group density as an empirical histogram. *Educational and Psychological Measurement*, 68, 571-586.

Zwick, R. (2010). The investigation of differential item functioning in adaptive test. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 331–352). New York: Springer.

Zwick, R., Thayer, D. T. & Wingersky, M. (1993) *A simulation study of methods for assessing differential item functioning in computer-adaptive tests*. (ETS Research Report 93-11). Princeton, NJ: Educationl Testing Service.

Zwick, R., Thayer, D. T.& Wingersky, M. (1994a) A simulation study of methods for assessing differential item functioning in computerized adaptive tests. *Applied Psychological Measurement*, 18, 121–140.

Zwick, R., Thayer, D. T. & Wingersky, M. (1994b) *DIF analysis for pretest items in computeradaptive testing* (ETS Research Report 94-33). Princeton, NJ: Educational Testing Service.

Zwick, R., Thayer, D. T. & Wingersky, M. (1995). Effect of Rasch calibration on ability and DIFestimation in computer-adaptive tests. *Journal of Educational Measurement*, 32, 341–363.

(Manuscript received: 7 January 2014; accepted: 27 March 2014)