The Effect Size for Simultaneous Item Bias Test

Zhushan Li
Boston College, Campion Hall 336D, 140 Commonwealth Ave, Chestnut Hill, MA 02467

Abstract

Differential item functioning (DIF) occurs when people from different groups with the same level of latent trait (ability/skills) have a different probability of responding to an item or a bundle of items in a questionnaire or test. DIF detection is an important step in the evaluation of the measurement bias in an instrument. Simultaneous Item Bias Test (SIBTEST, Shealy & Stout 1993) is a popular DIF detection method which can handle both dichotomous and polytomous items, and DIF in a single item and in a bundle of items. In this paper, we focus on the effect size measure as defined SIBTEST, and derive the formulas for the effect size under the IRT models. The relationship between the SIBTEST effect size and other popular DIF effect size measures are discussed. The correctness of the formula is confirmed by simulation studies.

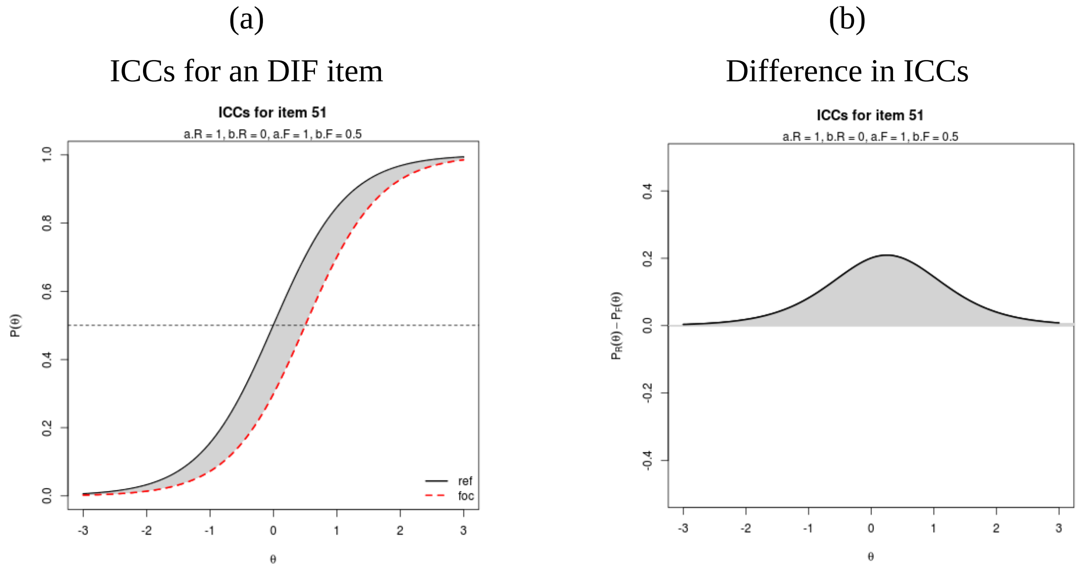
Key Words: DIF, effect size, item response theory, measurement bias

1. Introduction

In psychological and educational measurement, a fair instrument should not contain items that favor a specific group of people over other groups. In item response theory (IRT), the response characteristic for an item is described by the item characteristic curve (ICC) / item response function (IRF), which describes the relationship between the probabilities of getting the specific response (e.g. correct response for a binary item with possible correct/incorrect response). When the ICC for an item is different for people who come from different groups, then the item is biased and we call the function is a differential item functioning (DIF) item (Figure 1).

Methods have developed for DIF detection. Popular methods include Mantel-Haenszel test (Holland & Thayer, 1988; Mantel & Haenszel, 1959), logistic regression test (Swaminathan & Rogers, 1990), and the Simultaneous Item Bias Test (SIBTEST, Shealy & Stout 1993). In this paper, we will focus on the SIBTEST and provide details on the effect size in this procedure under the commonly used IRT model.

An important aspect of any statistical procedure is the power, which is the probability of making the decision to reject the null hypothesis when the alternative hypothesis is true. The power is related to the effect size, the sample size, and the significance level (Cohen, 1988). In practice, both the statistical significance (as describe by the p-value) and effect size should be considered in the decision process. While a statistical significance rule of $p < .05$ is nearly universally adopted, when the sample size is large, it is easy to get statistical significance results that may not be practically meaningful. An effect size describes the quantitative measure of a difference and is not subject to effect of the sample size, so it is used to describe the practical difference. In DIF detection, a decision on whether an item is a DIF item should be made upon both the statistical significance and the effect size. For example, in ETS classification scheme, DIF effects in an item are classified as negligible (Class A, |MH D-DIF| < 1 or the MH D-DIF is not significant at the level 0.05), moderate (Class B, 1 ≤ |MH D-DIF| < 1.5), and large (Class C, |MH D-DIF| ≥ 1.5 and the MH D-DIF is significantly greater than 1 in absolute value at the level 0.05), where MH D-DIF is the effect size measure used in MH test (Dorans & Holland, 1993). Similarly, one would argue that similar rules be used for other DIF procedures, including the SIBTEST.


(a)

ICCs for an DIF item

(b)

Difference in ICCs

Figure 1: A demonstration of differential item functioning in an item

2. SIBTEST and its effect size $\boldsymbol{\beta}$

Consider a test of length I with the first m items being the items with no DIF, and the remaining $I - m$ items being the studied items being suspected of DIF. Let U_i be the score for item i, $X = \sum_{i=1}^{m} U_i$ be the total score for the anchor items, and $Y = \sum_{i=m+1}^{I} U_i$ be the total score for the studied items. Let $\bar{Y}_{gk}$ be the average score on the studied items for all group g ($g = R$ or F; R for the reference group, and F for the focal group) examinees for which $X = k$.

The SIBTEST statistic is defined by the weighted sum of the local group differences by

$$\hat{\beta} = \sum_{k=0}^{m} p_k d_k = \sum_{k=0}^{m} p_k(\bar{Y}_{Rk} - \bar{Y}_{Fk})$$

where $p_k = N_k/N$ is the proportion of examinees (from the reference and focal groups pooled together) getting score $X = k$ ($p_{Fk} = N_{Fk}/N_F$ or $p_{Rk} = N_{Rk}/N_F$ can also be used in the place of p_k). In the formula

$$d_k = \bar{Y}_{Rk} - \bar{Y}_{Fk}, \;\; k = 0, \dots, m,$$

is the group difference on the studied items among examines with the same observed matching score on the anchor items. If the studied items have no DIF, one would expect $d_k \approx 0$.

The SIBTEST statistic is defined as the standardized version of $\hat{\beta}$ is defined by

$$B = \frac{\hat{\beta}}{\hat{\sigma}(\hat{\beta})},$$

where

$$\hat{\sigma}(\hat{\beta}) = \left[\sum_{k=0}^{m} p_k^2 \left(\frac{\hat{\sigma}^2(Y|k,R)}{N_{Rk}} + \frac{\hat{\sigma}^2(Y|k,F)}{N_{Fk}}\right)\right]^{\frac{1}{2}}$$

is the standard error of the estimator $\hat{\beta}$, and $\hat{\sigma}^2(Y|k,g)$ is the sample variance of the studied item scores for examinees in group g (*R* or *F*) with matching score $X = k$.

Under the observed-score DIF null hypothesis that $E_R(Y|X) = E_F(Y|X)$ for all X, B follows an asymptotic distribution of $N(0,1)$. a modified SIBTEST test statistic by a regression correction method was proposed in Shealy and Stout (1993); and later a more sophisticated regression correction method was proposed by Jiang and Stout (1998). By using the regression correction, a corrected version for d_k is given by $d_k^* = \bar{Y}_{Rk}^* - \bar{Y}_{Fk}^*$, $k = 0, \dots, m$, where $\bar{Y}_{gk}^*$ is the corrected average score for examinees matched on true score instead of observed score. The modified SIBTEST is given by

$$B^* = \frac{\hat{\beta}^*}{\hat{\sigma}(\hat{\beta})},$$

where $\hat{\beta}^*$ is the corrected version of $\hat{\beta}$. Under the latent-variable DIF null hypothesis that $E_R(Y|\theta) = E_F(Y|\theta)$, B^* has an asymptotic distribution of *N*(0,1).

If the test length is large and the number of examinees is large, $\hat{\beta}^*$ tends to the population parameter

$$\beta = \int_{-\infty}^{\infty} [E_R(Y|\theta) - E_F(Y|\theta)]\ f(\theta) d\theta,$$

where $f(\theta)$ is the density function for the population latent trait distribution. β is the effect size measure in the SIBTEST. And it can be interpreted as the weighted average score difference for the studied item (or a bundle of studied items) between the reference and focal groups. For a single dichotomous studied item, β is the weighted average difference in probability of getting the correct response to the item. So β is a “difference in probability or difference in score” measure.

3. Formula for effect size β under IRT models

Consider the item response function for the IRT 3PL model

$$P_{3PL}(\theta; a, b, c) = c + (1 - c)\frac{\exp\big(Da(\theta - b)\big)}{1 + \exp\big(Da(\theta - b)\big)},$$

where θ is the latent trait a is the discrimination parameter, *b* is the difficutly parameter, *c* is the guessing parameter, and *D* is a scale factor. In the IRT literature, $D = 1.7$ is often used for the 3PL and 2PL models, because the item response curve is very close to that of the normal ogive model with mean *b* and standard deviation $1/a$ (Birnbaum, 1968; Camilli, 1994). In this paper, we will use *D* = 1.7 for the 3PL and 2PL models. So, 2PL model is obtained by setting $c = 0$, and the 1PL/Rasch model is given by setting $c = 0$ and $a = 1$.

In this section, we will derive the approximate formula for effect size β. We assume that the 3PL model is satisfied with restrictions $c_R = c_F = c$ and $a_R = a_F = a$, and the assumption $f_R(\theta) = f_F(\theta) = f(\theta) \sim N(\mu, \sigma^2)$. In other words, only difference between the two IRF's between the reference and focal groups are in difficulty parameter.

The item response functions for the reference and focal groups are $P_R(\theta) = P_{3PL}(\theta; a, b_R, c) = c + (1-c)P_{2PL}(\theta; a, b_R)$, and $P_F(\theta) = P_{3PL}(\theta; a, b_F, c) = c + (1-c)P_{2PL}(\theta; a, b_F)$, where P_{2PL} is given by

$$P_{2PL}(\theta; a, b) = \frac{\exp\big(1.7a(\theta - b)\big)}{1 + \exp\big(1.7a(\theta - b)\big)}.$$

Therefore, by the linear approximation and the logistic-normal approximation,

$$\begin{aligned} P_R(\theta) - P_F(\theta) &= (1-c)[P_{2PL}(\theta; a, b_R) - P_{2PL}(\theta; a, b_F)] \\ &\approx (1-c)(b_F - b_R)a\phi\big(a(\theta - b_R)\big). \end{aligned}$$

And then according to the formulas proved in Appendix,

$$\begin{aligned} \beta &= \int_{-\infty}^{\infty} [P_R(\theta) - P_F(\theta)] f(\theta) d\theta \\ &\approx (1-c)(b_F - b_R) \int_{-\infty}^{\infty} a\phi\big(a(\theta - b)\big) \left(\frac{1}{\sigma}\right) \phi\left(\frac{\theta - \mu}{\sigma}\right) d\theta \\ &= (1-c)(b_F - b_R) \frac{1}{\sqrt{a^{-2} + \sigma^2}} \phi\left(\frac{b - \mu}{\sqrt{a^{-2} + \sigma^2}}\right) \\ &= (1-c)\Delta b \frac{1}{\sqrt{a^{-2} + \sigma^2}} \phi\left(\frac{b - \mu}{\sqrt{a^{-2} + \sigma^2}}\right), \end{aligned}$$

which is the closed-form approximate formula for the effect size for IRT 3PL model. From this formula one can see that the effect size β is approximately proportional to $\Delta b = b_F - b_R$, which is the difference in difficulty, note that Δb itself is often used as an effect size for DIF. The guessing parameter c will reduce the effect size β by a factor of $(1 - c)$. Note that the second half of the approximate formula is the density function of the Normal distribution evaluated at $(b - \mu)/\sqrt{a^{-2} + \sigma^2}$, which obtain maximum when $b = \mu$, i.e. if the item difficulty is the same as the population latent trait mean, then β is the largest.

4. Effect size β and other DIF effect size measures

As we have seen, the SIBTEST effect size β can be interpreted as the weighted average score difference for the studied item (or a bundle of studied items) between the reference and focal groups. For a single dichotomous studied item, β is the weighted average difference in probability of getting the correct response to the item. So β is a "difference in probability or difference in score" measure. We now discuss its relationship to other popular effect size measures used in other DIF testing procedures.

4.1 Raju's area between IRF curves

A natural global measure for DIF is the area between the two IRF curves

$$\int_{-\infty}^{\infty} [P_R(\theta) - P_F(\theta)]\, d\theta\,.$$

Raju (1988) gives a general formula for the area between two IRF's under the 3PL models. Under the restriction that $c_R = c_F = c$, and allowing either $a_R \neq a_F$ or $a_R = a_F$, the area is $(1-c)(b_F - b_R)$. Therefore, under the 2PL model, assuming that $a_R = a_F = a$, the local DIF effect $[P_R(\theta) - P_F(\theta)]$ has an area under the curve equal to

$\int_{-\infty}^{\infty}[P_R(\theta)-P_F(\theta)]d\theta = b_F - b_R$, regardless the discrimination parameter a . Furthermore, under the 3PL model, assuming that the only difference in IRF between the reference group and the focal group is the difficulty parameter, i.e., $a_R = a_F = a$ and $c_R = c_F = c$, and $b_R \neq b_F$, the integral now becomes $\int_{-\infty}^{\infty}[P_R(\theta)-P_F(\theta)]d\theta = (1-c)(b_F - b_R)$. Therefore, guessing will reduce the DIF effect size.

One problem of using the area-between-curves measure is that it is not always finite. For example, in the case of 3PL model when $c_R \neq c_F$, the area between the curves is infinity and, thus, undefined. The area-between-curves measure gives equal weight for θ along the entire real line, ignoring the fact that population ability is more likely to occur in some regions than in others. Therefore, one sensible alternative measure is the area between the IRFs weighted by some latent trait distribution density

$$\int_{-\infty}^{\infty}[P_R(\theta)-P_F(\theta)]\, f^*(\theta)d\theta \,,$$

which is always finite. There are different choices for $f^*(\theta)$: (1) the focal group latent trait distribution density, $f_F(\theta)$, (2) the reference group latent trait distribution density, $f_R(\theta)$, or (3) pooled density function $\gamma_F f_F(\theta) + (1-\gamma_F)f_R(\theta)$. Standardized p-difference (Dorans & Kulick, 1986) and the original SIBTEST paper (Shealy & Stout, 1993) use choice (1), and the current SIBTEST software provides options for all three. In this paper, we discuss the β measure under choice (3).

4.2 Mantel-Haenszel Odds Ratio

Another popular effect size measure of DIF is the odds ratio statistic associated with Mantel-Haenszel DIF procedure (Holland & Thayer, 1988). The MH odds ratio statistic can be seen as an estimate of some average of the local odds ratio $\alpha(\theta) = [\, P_R(\theta)Q_F(\theta)]/[\, P_F(\theta)Q_R(\theta)]$. Under the 2PL model with common discrimination parameter, the logarithm of the local odds ratio is $\log\alpha(\theta) = 1.7a(b_F - b_R) = 1.7a\Delta b$, so it is called uniform DIF because the DIF effect size is constant for all θ. In this case, the global log MH odds ratio effect size measure is the same constant: $\log\alpha_{MH} = 1.7a\Delta b$. For the 2PL model with different discrimination parameter, and 3PL model, $\alpha(\theta)$ is no longer uniform with respect to θ, and the global effect size measure $\log\alpha_{MH}$ is rather complicated (See Roussos, Schnipke, & Pashley, 1999 for discussions on this effect size measure for the 3PL model).

4.3 Delta-b measure

One can also directly adopt a “delta-b” measure: $\Delta\, b = b_R - b_F$ as an effect size measure for DIF. The log MH odds ratio and delta-b can be classified as “difference in difficulty” measures. A “difference in probability” measure and a “difference in difficulty” measure are connected but not totally aligned and there is no obvious reason to believe one is better than the other. For example, in the early era of the development of DIF methods, both MH and standardized p-DIF were adopted as standard DIF tools in ETS practice (Dorans & Holland, 1993). In this paper, we give examples for both $\Delta\, b$ and β. The two effect sizes are on different scales and there is no strict one-to-one relationship between the two measures because the relationship also involves the other factors.

4.4 Small, Medium, and Large Effect Size

In practice, one often would like to use qualitative adjectives such as “small”, “medium”, or “large” to describe an effect size and to help guide decision making. ETS uses a classification system to classify DIF effect sizes into three categories. It is based on the MH D-DIF index, which is equal to -2.35 times the log MH odds ratio. DIF effects are classified as negligible (Class A, |MH D-DIF| < 1 or the MH D-DIF is not significant at the level 0.05), moderate (Class B, 1 ≤ |MH D-DIF| $<$ 1.5), and large (Class C, |MH D-DIF| ≥ 1.5 and the MH D-DIF is significantly greater than 1 in absolute value at the level 0.05) (Dorans & Holland, 1993). Based on the ETS classification scheme, a value of Δb =0.426 (equivalent MH D-DIF = 1) is considered a medium effect, and Δb =0.638 (equivalent MH D-DIF = 1.5) is considered a large effect under the Rasch model (Paek & Wilson, 2011). For SIBTEST effect size β, there are no single direct one-to-one translation of these two Δb values into β values that applies to all possible situations because of other factors also involved the relationship. However, if we use a relationship we derived earlier, $\beta \approx 0.18 \vartriangle b$ under a typical situation (Rasch model, b_R within 1 *SD* from the population mean ability, and $\sigma = 1$), also see β values in Table 2 later in this paper, then we would consider β = 0.08 as a medium effect, and β = 0.12 as a large effect. In the literature, another set of commonly used values are $\beta = 0.05$ as a medium effect and $\beta = 0.10$ as a large effect, which were first suggested by Dorans (1989).

5. Simulation study

In the simulation study, we consider a test of 100 items. Items 1 to 50 are anchor items: the validated items that are known to have no DIF. Items 51 to 100 are DIF items. All the items follow the IRT 2PL model. For each item, the discrimination parameter value a (here we assume $a_R = a_F = a$) is chosen by a positive random draw from a normal distribution with mean 1.2 and variance 0.1 (i.e., *SD* = 0.32); the reference group difficulty parameter b_R is drawn from a normal distribution with mean 0 and standard deviation 1. For anchor items 1 to 50, the focal group difficulty parameter b_F is set equal to b_R. For DIF items 51 to 100, b_F is set equal to $b_R + \Delta b$, where $\Delta b = 0.1$ for items 51-60, 0.2 for items 61-70, 0.3 for items 71-80, 0.4 for items 81-90, and 0.5 for items 91-100. The item parameter values were recorded and saved for later calculation of theoretical power. The same items were used in 10,000 replicated simulation runs. In each run, a sample of $N_R = 1000$ examinees from the reference group and $N_F = 1000$ examinees from the focal group were simulated by drawing their ability parameter θ from a $N(0, 1)$ distribution. Then for each examinee in the reference group, the dichotomous response to each of the 100 items was simulated based on the 2PL model with discrimination a and difficulty b_R; and for each examinee in the focal group, based on the 2PL model with discrimination a and difficulty b_F. The simulated dataset was then analyzed by the SIBTEST program for detecting DIF in each item. In the SIBTEST analysis, the total score of the first 50 anchor items was used as the matching variable to stratify the examinees, and the SIBTEST procedure with a two-sided alternative hypothesis was conducted on each of the 100 items. For each item, the following estimates were produced by the SIBTEST program and recorded: $\hat{\beta}$, SE, B, pvalue, and the decision of the test (reject H_0 if $p < .05$, and do not reject H_0 otherwise).

Based on 10,000 replicated simulation runs, the following quantities were summarized for each item:

- the mean of $\hat{\beta}$, which should be close to the true value of the effect size β;
- the standard deviation of $\hat{\beta}$, which is the Monte Carlo SE;
- the rejection rate (= the number of times rejecting H_0 / 10,000).

For the anchor items with no DIF, the rejection rate is an estimate of the Type I error rate, which is expected to be close to 0.05. For the DIF items, the rejection rate is an estimate of the true power of the SIBTEST procedure, which we will use to check the validity of the power formula. The rejection rates for the first 50 non-DIF items are close to 0.05, and most of the values (47 out of 50; 94%) are located within the 95% confidence band given by $0.05 \pm 1.96\sigma_M$, where σ_M is the Monte Carlo error and its value is calculated by $\sqrt{(0.05 * 0.95)/10{,}000} = 0.0022$. This indicates that the Type I error rate of SIBTEST is well controlled at 0.05. The rejection rates of the last 50 DIF items has a general trend of increasing as Δb increases, with the variability from this trend observed be due to the fact that the effect size is affect by other factors such as item discrimination a, and thus the power is not solely determined by Δb.

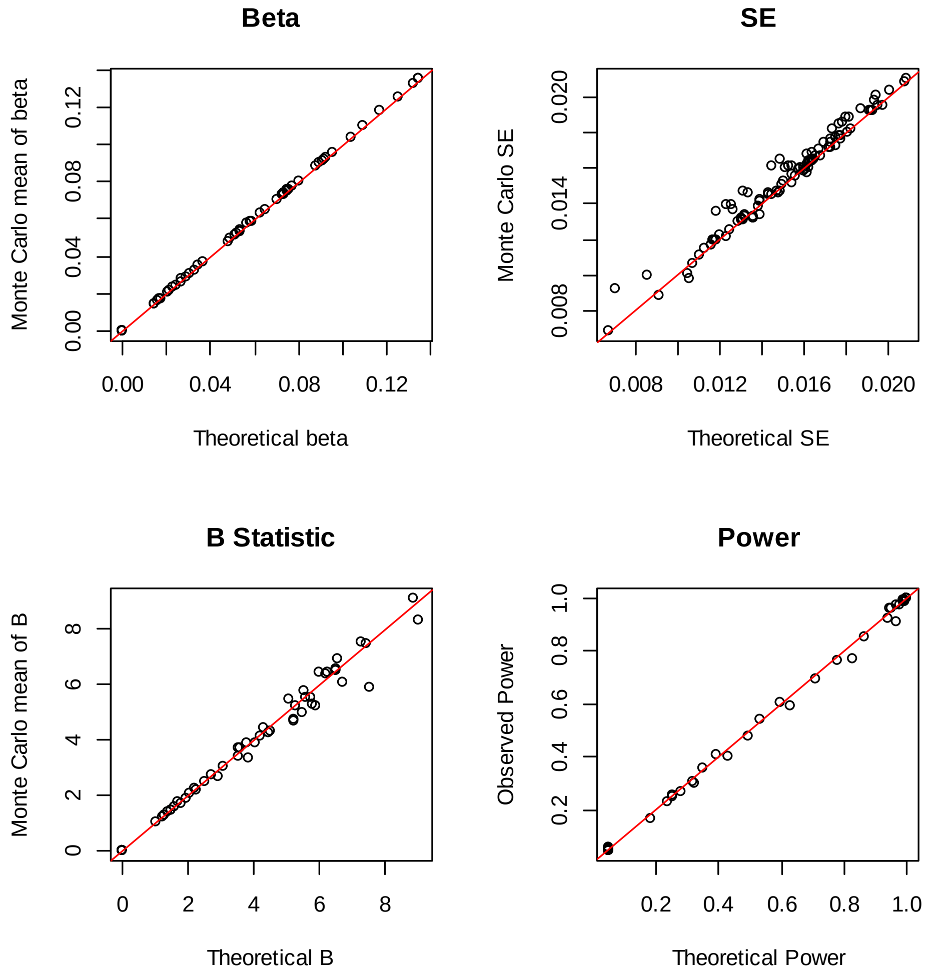


Figure 2. Comparison of theoretical and observed values from simulation study.

Now let us compare the results obtained from the simulation data with the values based on the formula derived in this paper. Figure 2 shows scatter plots comparing four key quantities in the SIBTEST power formula: β, SE, B, and power. In each plot, the y-axis represents the values observed in the simulation data, which are considered as very close to "truth" based on the law of large numbers; and the x-axis represents the "theoretical" values calculated from the formula derived in this paper, using the parameters used for simulating the data. We can see that in all the scatter plots, the points are close to the 45-degree line, which means that the theoretical results agree very well with those observed in the simulation results. The root mean square difference (RMSD) for the four quantities are: $RMSD(\beta) = 0.00025$, $RMSD(SE) = 0.0006$, $RSMD(B) = 0.258$, $RSMD(\text{power}) = 0.011$.

References

Birnbaum, A. (1968). Some latent train models and their use in inferring an examinee's ability. In F. M. Lord, M. R. Novick & A. Birnbaum (Eds.), *Statistical theories of mental test scores* (pp. 395-479)

Camilli, G. (1994). Origin of the scaling constant d=1.7 in item response theory. *Journal of Educational and Behavioral Statistics, 19*(3), 293-295.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*: Routledge Academic.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: standardization and the Mantel-Haenszel method. *Applied Measurement in Education, 2*(3), 217-233.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning: Theory and Practice* (pp. 35-66). Hillsdale, NJ: 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*(4), 355-368.

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

Jiang, H., & Stout, W. (1998). Improved Type I error control and reduced estimation bias for DIF detection using SIBTEST. *Journal of Educational Statistics, 23*(4), 291-322.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*(4), 719-748.

Paek, I., & Wilson, M. (2011). Formulating the Rasch differential item functioning model under the marginal maximum likelihood estimation context and its comparison with Mantel-Haenszel procedure in short test and small sample conditions. *Educational and Psychological Measurement, 71*(6), 1023-1046.

Roussos, L. A., Schnipke, D. L., & Pashley, P. J. (1999). A generalized formula for the Mantel-Haenszel differential item functioning parameter. *Journal of Educational and Behavioral Statistics, 24*(3), 293-322.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*(2), 159-194.

Swaminathan, H., & Rogers, H. J. (1990). Detecting Differential Item Functioning Using Logistic Regression Procedures. *Journal of Educational Measurement, 27*(4), 361-370.

Appendix

As prerequisite for proving the formulas in the main text, let us prove the following two equations: assuming $a > 0$ and $\sigma > 0$,

$$\int_{-\infty}^{\infty} \frac{1}{a}\phi\left(\frac{x-b}{a}\right)\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)dx = \frac{1}{\sqrt{a^2+\sigma^2}}\phi\left(\frac{b-\mu}{\sqrt{a^2+\sigma^2}}\right),$$

and

$$\int_{-\infty}^{\infty} \Phi\left(\frac{x-b}{a}\right)\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)dx = 1-\Phi\left(\frac{b-\mu}{\sqrt{a^2+\sigma^2}}\right).$$

where $\phi(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$ is the density function of the standard normal distribution, and $\Phi(x) = \int_{-\infty}^{x}\phi(s)ds$ is the distribution function of the standard normal distribution.

The proof of the first equation is

$$\begin{aligned}
&\int_{-\infty}^{\infty} \frac{1}{a}\phi\left(\frac{x-b}{a}\right)\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}a}\exp\left(-\frac{(x-b)^2}{2a^2}\right)\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx \\
&= \frac{1}{2\pi a\sigma}\int_{-\infty}^{\infty}\exp\left(-\frac{a^2+\sigma^2}{2a^2\sigma^2}\left[x^2-\frac{2(b\sigma^2+\mu a^2)}{a^2+\sigma^2}x+\frac{b^2\sigma^2+\mu^2a^2}{a^2+\sigma^2}\right]\right)dx \\
&= \frac{1}{2\pi a\sigma}\int_{-\infty}^{\infty}\exp\left(-\frac{a^2+\sigma^2}{2a^2\sigma^2}\left[x-\frac{(b\sigma^2+\mu a^2)}{a^2+\sigma^2}\right]^2-\frac{(b-\mu)^2}{2(a^2+\sigma^2)}\right)dx \\
&= \frac{1}{2\pi a\sigma}\exp\left(-\frac{(b-\mu)^2}{2(a^2+\sigma^2)}\right)\int_{-\infty}^{\infty}\exp\left(-\frac{a^2+\sigma^2}{2a^2\sigma^2}\left[x-\frac{(b\sigma^2+\mu a^2)}{a^2+\sigma^2}\right]^2\right)dx \\
&= \frac{1}{2\pi a\sigma}\exp\left(-\frac{(b-\mu)^2}{2(a^2+\sigma^2)}\right)\sqrt{2\pi\frac{a^2\sigma^2}{a^2+\sigma^2}} \\
&= \frac{1}{\sqrt{2\pi(a^2+\sigma^2)}}\exp\left(-\frac{(b-\mu)^2}{2(a^2+\sigma^2)}\right) \\
&= \frac{1}{\sqrt{a^2+\sigma^2}}\phi\left(\frac{b-\mu}{\sqrt{a^2+\sigma^2}}\right).
\end{aligned}$$

Take the partial derivative of $I = \int_{-\infty}^{\infty}\Phi\left(\frac{x-b}{a}\right)\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)dx$ with respect to b, and then apply first equation, we have

$$\begin{aligned}
\frac{\partial I}{\partial b} &= -\int_{-\infty}^{\infty}\frac{1}{a}\phi\left(\frac{x-b}{a}\right)\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)dx \\
&= -\frac{1}{\sqrt{a^2+\sigma^2}}\phi\left(\frac{b-\mu}{\sqrt{a^2+\sigma^2}}\right)
\end{aligned}$$

Therefore

$$I = \int -\frac{1}{\sqrt{a^2+\sigma^2}}\phi\left(\frac{b-\mu}{\sqrt{a^2+\sigma^2}}\right)db = -\Phi\left(\frac{b-\mu}{\sqrt{a^2+\sigma^2}}\right)+C$$

Since $b = -\infty$, the integral $\int_{-\infty}^{\infty}\Phi\left(\frac{x-b}{a}\right)\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)dx = 1$, solving above equation we get the constant $C = 1$.