A Power Formula for the Mantel–Haenszel Test for Differential Item Functioning

Applied Psychological Measurement

2015, Vol. 39(5) 373–388

© The Author(s) 2015

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0146621614568805

apm.sagepub.com

Zhushan Li¹

Abstract

The asymptotic power of the Mantel–Haenszel (MH) test for the differential item function (DIF) is derived. The formula describes the behavior of the power when the number of items is large, so that the measured latent trait can be considered as the matching variable in the MH test. As shown in the derived formula, the power is related to the sample size, effect size of DIF, the item response function (IRF), and the distribution of the latent trait in the reference and the focal groups. The formula provides an approximation of the power of the MH test in practice and thus provides a guideline for DIF detection in practice. It also suggests analytical explanations of the behavior of the MH test as observed in many previous simulation studies. Based on the formula, this study shows how to conduct the sample size calculation. The power of MH test under some practical models such as the two-parameter logistic (2PL) and three-parameter logistic (3PL) item response theory (IRT) models is discussed.

Keywords

DIF, Mantel–Haenszel test, item response theory, power

Differential item function (DIF) occurs when the chance to get correct answer to an item differs for people having the same level of ability the item is designed to measure but from different populations (Holland & Wainer, 1993; Lord, 1980; Millsap & Everson, 1993). DIF detection and related topics have been continuously in active research (Cheng, Chen, Qian, & Chang, 2013; Finch, 2012; Wang, Tay, & Drasgow, 2013). The Mantel–Haenszel (MH) test for DIF (Holland & Thayer, 1988) is one of the most popular methods to detect DIF. The method is nonparametric in the sense that it does not make assumptions on the specific form of the item response function (IRF) and the underlying latent trait distribution. It is the uniform most powerful (UMP) test for detecting uniform DIF effect (Birch, 1964) and the test statistic is very easy to calculate. With these advantages, the MH test for DIF has been widely adopted in practice.

Since the introduction of the MH test in DIF research, there have been many studies on the power of the MH test and through these studies, many factors that may influence the power became well understood (Donoghue, Holland, & Thayer, 1993; Paek, 2010; Roussos & Stout, 1996; Swaminathan & Rogers, 1990; Uttaro & Millsap, 1994). However, most of those studies

¹Boston College, Chestnut Hill, MA, USA

Corresponding Author:

Zhushan Li, Assistant Professor, Boston College, Campion Hall, Room 336D, 140 Commonwealth Avenue, Chestnut Hill, MA 02467, USA.

Email: zhushan.li@bc.edu

were simulation studies, where the power of the MH test was estimated by the rejection rate of the MH test applied on the simulated data sets generated under specific designs of the underlying model and parameters. And there is still a lack of a theoretical formula for the power of the MH DIF test. Simulation studies are very useful in identifying factors that influence the power of the MH test, but they provide very little explanation about why the power is influenced by those factors. Without a clear understanding of the underlying mechanism, it is difficult to say whether the knowledge learned from particular simulation studies still holds when facing a new test situation. One may conduct new simulation studies specifically tailored to that new situation, but practitioners often lack the time and resources to conduct extensive large-scale simulation studies. However, a power formula would provide insights on the factors in the formula, and its applicability is general, and a formula is straightforward to use in practice.

In practice, a DIF analysis is often conducted only when sufficient sample sizes are reached; otherwise, it would be a waste of the resources due to the lack of the power. For example, at ETS, the current requirements for DIF analyses at the test assembly phase are at least 200 people in a smaller group and at least 500 in total (200/500 rule); and at post-administration and prescore reporting phases, the requirements are at least 300 in a smaller group and 700 in total (300/700 rule; Zwick, 2012). One can argue that such rules are rather arbitrary. Generally speaking, the power is an intrinsic property of a statistical test; and it is related to the effect size, the sample size, and the significance level (Cohen, 1988). For DIF analysis, there are many factors that affect the power of the test, and there is no single rule that fits all possible situations. Although in other areas of behavioral sciences where t tests, ANOVAs, and so on are conducted, power analyses and sample size calculations are routinely done or recommended at the research planning stage, for DIF studies, they are seldom conducted mainly due to the lack of tools available to practitioners.

This article aims at theoretically deriving a general formula for the power function of the MH test for DIF detection, and applying the formula to the Rasch, item response theory (IRT) two-parameter logistic (2PL) and three-parameter logistic (3PL) models. Power formulas have been developed in recent articles for SIBTEST (Li, 2014b) and logistic regression DIF tests (Li, 2014a). With the power function formula, it is possible to not only identify but also explain the factors that influence the MH test, and conduct power and sample size calculations for the experimental design of DIF studies.

DIF and MH Test

Consider a test consisting of I items, where the response to the items is binary and coded as $Y_i = 1$ if the answer to the item i is correct, and $Y_i = 0$ if incorrect. The test-takers are from two population groups, one is the focal group as coded by $G = F$, and the other is the reference group as coded by $G = R$. The problem is to test whether an item of interest (the studied item) is a DIF item. The remaining $I - 1$ items are usually considered as non-DIF items (anchor items). An item is a non-DIF item if the probability of getting the studied item correct is the same regardless of the test-takers' group membership after the ability is matched; otherwise, the item is a DIF item. By matching on the total score V , the examinees are grouped into $I + 1$ strata by the scores $0, 1, \dots, I$. Within each stratum, a 2×2 cross table between the group membership and the response to the studied item can be constructed (Table 1).

In the table, A_j , B_j , C_j , and D_j are the observed counts of examinees in each cell in the j th stratum; $N_j = A_j + B_j + C_j + D_j$ is the number of examinees in the stratum; $n_{Rj} = A_j + B_j$ and $n_{Fj} = C_j + D_j$ are the numbers of examinees in the reference and focal groups in the stratum, respectively; and $m_{1j} = A_j + C_j$ and $m_{0j} = B_j + D_j$ are the numbers of examinees who answer the studied item correctly and incorrectly, respectively. The null hypothesis of no DIF states that

Table 1. 2×2 Table for the j th Matched Set of Members of Reference and Focal Groups.

Group	Reference ($G=R$) Focal ($G=F$)	Response to the studied item		n_{Rj} n_{Fj} N_j
		Correct ($Y=1$)	Incorrect ($Y=0$)	
		A_j C_j m_{1j}	B_j D_j m_{0j}	

conditional on the matching variable, there is no association between the group membership and the response to the studied item.

The MH chi-square test statistic (Mantel & Haenszel, 1959) is given by,

$$MH - \chi_C^2 = \frac{\left(\left| \sum_j A_j - \sum_j E_{A_j} \right| - 0.5 \right)^2}{\sum_j V_{A_j}}, \quad (1)$$

where $E_{A_j} = (A_j + B_j)(A_j + C_j)/N_j$, and

$$V_{A_j} = \frac{n_{Rj}n_{Fj}m_{1j}m_{0j}}{N_j^2(N_j - 1)},$$

are the estimates of the mean and the variance, respectively, of the cell count A_j under the null hypothesis and the hypergeometric sampling model (Agresti, 2002). Under the null hypothesis, the MH chi-square statistic has a chi-square distribution with degree of freedom 1. The term “−0.5” in the MH chi-square test statistic is a continuity correction, which is supposed to improve accuracy of Type I error (Holland & Thayer, 1988). It is possible to define the MH chi-square statistic without the continuity correction as follows:

$$MH - \chi^2 = \frac{\left(\sum_j A_j - \sum_j E_{A_j} \right)^2}{\sum_j V_{A_j}}. \quad (2)$$

Previous simulation studies suggest that the MH chi-square with the continuity correction tends to be more conservative than the MH chi-square without the continuity correction (Paek, 2010). For example, at the level of 0.05, it is often observed in simulated data with smaller test length that the actual Type I error rate for the MH chi-square test with continuity correction is about 0.03 to 0.04, whereas for the MH-chi-square test without continuity correction on the same data set, it is about 0.05.

The MH chi-square statistics in Equations 1 or 2 do not directly give the direction of the DIF effects if they are significant. Holland and Thayer (1988) use the following formula to estimate the common odds ratio:

$$\hat{\alpha}_{MH} = \frac{R}{S} = \frac{\sum_j R_j}{\sum_j S_j} = \frac{\sum_j \frac{A_j D_j}{N_j}}{\sum_j \frac{B_j C_j}{N_j}}. \quad (3)$$

The MH log odds ratio estimate $\hat{\psi}_{MH} = \log \hat{\alpha}_{MH}$ has an asymptotic normal distribution around the true value $\psi_{MH} = \log \alpha_{MH}$. A formula for the variance of the MH log odds ratio estimate that was derived by Phillips and Holland (1987) is given by

$$\widehat{\text{Var}}_{PH}(\hat{\psi}_{MH}) = \frac{1}{2R^2} \sum_j N_j^{-2} (A_j D_j + \hat{\alpha}_{MH} B_j C_j) [A_j + D_j + \hat{\alpha}_{MH} (B_j + C_j)]. \quad (4)$$

Therefore, a MH-Z statistic can be defined by dividing the log odds ratio estimate by its standard error:

$$\text{MH-Z} = \frac{\hat{\psi}_{MH}}{\sqrt{\widehat{\text{Var}}(\hat{\psi}_{MH})}}. \quad (5)$$

The MH-Z statistic is asymptotically equivalent to the $\text{MH-}\chi^2$ statistic. Under the null hypothesis, the MH-Z statistic has an asymptotic standard normal distribution, so $(\text{MH-Z})^2$ has an asymptotic distribution of chi-square with 1 degree of freedom. Under the alternative hypothesis, $(\text{MH-Z})^2$ and $\text{MH-}\chi^2$ have the same asymptotic distribution of non-central chi-square with 1 degree of freedom. The proof of this equivalence is shown in the online appendix. In this article, the asymptotic power formula is derived based on the MH-Z statistic. Because of the asymptotic equivalence, the same formula is also applicable for the $\text{MH-}\chi^2$ statistic. Later in simulation studies, it is also shown that the rejection rates of the test based on MH-Z statistic is very similar to those of the $\text{MH-}\chi^2$ test.

Asymptotic Formula for $\hat{\alpha}_{MH}$ and the Variance Estimate of $\log \hat{\alpha}_{MH}$

Let $P_R(\theta)$ and $P_F(\theta)$ be the IRF of the studied item for the reference group and the focal group, respectively: $P_R(\theta) = P(Y = 1 | \theta, G = R)$ and $P_F(\theta) = P(Y = 1 | \theta, G = F)$; and let $Q_R(\theta) = 1 - P_R(\theta)$ and $Q_F(\theta) = 1 - P_F(\theta)$.

Define the local odds ratio $\alpha(\theta)$ be the odds ratio given θ ,

$$\alpha(\theta) = \frac{P_R(\theta)Q_F(\theta)}{P_F(\theta)Q_R(\theta)}.$$

Roussos, Schnipke, and Pashley (1999) gave a general asymptotic formula for the MH odds ratio estimate $\hat{\alpha}_{MH}$. As test length tends to ∞ and the sample size tends to ∞ , matching on total score tends to matching on θ . Then, the MH odds ratio will tend to,

$$\hat{\alpha}_{MH} \rightarrow \alpha_{\infty} = \frac{\int P_F(\theta)Q_R(\theta) \frac{f_R(\theta)f_F(\theta)}{\gamma_F f_F(\theta) + \gamma_R f_R(\theta)} \alpha(\theta) d\theta}{\int P_F(\theta)Q_R(\theta) \frac{f_R(\theta)f_F(\theta)}{\gamma_F f_F(\theta) + \gamma_R f_R(\theta)} d\theta}, \quad (6)$$

where γ_R and γ_F are the proportion of reference group and focal group test-takers in their pooled population. Thus, $\gamma_R = P(G = R)$ and $\gamma_F = P(G = F) = 1 - \gamma_R$. Furthermore, $f_R(\theta)$ and $f_F(\theta)$ are the probability density functions (PDFs) for θ in the reference and focal group populations: $f_R(\theta) = p(\theta | G = R)$ and $f_F(\theta) = p(\theta | G = F)$.

To derive a power formula for the MH-Z statistics, a general formula is also necessary for the variance of the log odds ratio estimates under the null hypothesis, where $P_R(\theta) = P_F(\theta)$. Starting from the Phillips and Holland (1987) formula (Equation 4), under the null, $\hat{\alpha}_{MH} \rightarrow \alpha_{MH} = 1$, $\hat{\alpha}_j = A_j D_j / B_j C_j \rightarrow \alpha_{MH} = 1$, therefore,

$$\widehat{\text{Var}}_{\text{PH}}(\hat{\psi}_{\text{MH}}|H_0) \approx \frac{1}{2R^2} \sum_j 2S_j = \frac{S}{R^2} \approx \frac{1}{R} \rightarrow \frac{1}{n \int P_R(\theta) Q_R(\theta) f_R(\theta) f_F(\theta) \left(\frac{1}{p(\theta)}\right) \gamma_R \gamma_F d\theta}. \quad (7)$$

Deriving the Power Formula for MH Test

Under the null hypothesis, the MH-Z statistic follows the $N(0,1)$ distribution. So for the one-sided test, $H_0 : \psi = 0$ versus $H_1 : \psi > 0$, the null hypothesis is rejected if $\text{MH} - Z > z_\alpha$ for Type I error α , where the critical value, $z_\alpha = \Phi^{-1}(1 - \alpha)$ and Φ , is the distribution function for the standard normal distribution (e.g., for $\alpha = 0.05$, $z_\alpha = 1.645$). The power is the probability of rejecting the null under the alternative hypothesis. Under H_1 , MH-Z follows asymptotic distribution $N(\sqrt{n}\psi/\tau(\psi), 1)$. So the power is,

$$\text{power} = \Pr(\text{reject } H_0 | H_1) = \Phi\left(\frac{\sqrt{n}\psi}{\tau(\psi)} - z_\alpha\right).$$

For details on the theory on asymptotic power, see Lehmann (1999).

Therefore, the asymptotic power for the one-sided DIF test ($H_1 : \psi > 0$ or $H_1 : \psi < 0$) is given by

$$\begin{aligned} \text{power} &= \Phi\left(\frac{\sqrt{n}|\psi|}{\tau(\psi_0)} - z_\alpha\right) \\ &= \Phi\left(\sqrt{n}|\psi_{\text{MH}}| \left[\int_{-\infty}^{\infty} P_R(\theta) Q_R(\theta) \frac{f_R(\theta) f_F(\theta) \gamma_R \gamma_F}{\gamma_R f_R(\theta) + \gamma_F f_F(\theta)} d\theta \right]^{\frac{1}{2}} - z_\alpha\right). \end{aligned} \quad (8)$$

And for the two-sided DIF test ($H_1 : \psi \neq 0$), the power is given by

$$\begin{aligned} \text{power} &= \Phi\left(\frac{\sqrt{n}\psi}{\tau(\psi_0)} - z_{\alpha/2}\right) + \Phi\left(-\frac{\sqrt{n}\psi}{\tau(\psi_0)} - z_{\alpha/2}\right) \\ &= \Phi\left(\sqrt{n}\psi_{\text{MH}} \left[\int_{-\infty}^{\infty} P_R(\theta) Q_R(\theta) \frac{f_R(\theta) f_F(\theta) \gamma_R \gamma_F}{\gamma_R f_R(\theta) + \gamma_F f_F(\theta)} d\theta \right]^{\frac{1}{2}} - z_{\alpha/2}\right) \\ &\quad + \Phi\left(-\sqrt{n}\psi_{\text{MH}} \left[\int_{-\infty}^{\infty} P_R(\theta) Q_R(\theta) \frac{f_R(\theta) f_F(\theta) \gamma_R \gamma_F}{\gamma_R f_R(\theta) + \gamma_F f_F(\theta)} d\theta \right]^{\frac{1}{2}} - z_{\alpha/2}\right). \end{aligned} \quad (9)$$

The integrals in the power formulas can be evaluated by numeric integration methods. The author has used the numeric integration procedure in R (R Core Team, 2012) to obtain the numeric results in this article.

Factors That Influence the Power of the MH Test

From Equations 8 and 9, it can be shown that the power of the MH test when the number of items is large is related to

1. The sample size n —As the sample size increases, the power will increase.

2. The effect size $\psi_{MH} = \log \alpha_{MH}$ —Larger effect size results in larger power. In the case of uniform DIF (Rasch or 2PL), ψ_{MH} has a simple proportional relationship to $a(b_F - b_R)$. In the case of non-uniform DIF (3PL), Roussos et al. (1999) found that “for items with medium or high difficulty, the parameter is much smaller in absolute value than what is expected” in the case of uniform DIF.
3. The integral that is related to both the IRF and the distribution of the latent variable—The test will have higher power if the integral is larger.
4. The Type I error α —The power will increase if one is willing to increase the Type I error. However, in practice, the Type I error is usually fixed, for example, $\alpha = .05$.

A detailed discussion is presented on the components in the integral

$$\int_{-\infty}^{\infty} P_R(\theta) Q_R(\theta) \frac{f_R(\theta) f_F(\theta) \gamma_R \gamma_F}{\gamma_R f_R(\theta) + \gamma_F f_F(\theta)} d\theta,$$

and how they influence the power. The reciprocal of the integral is $\tau(\psi_0)^2$, and $\tau(\psi_0)/\sqrt{n}$ is the standard error of the log odds ratio estimate $\hat{\psi} = \log \hat{\alpha}_{MH}$ under the null hypothesis. The integrand is the product of two parts: the first part $P_R(\theta) Q_R(\theta)$ is only related to the IRF; and the second part, which is denoted by $H(\theta) = f_R(\theta) f_F(\theta) \gamma_R \gamma_F / [\gamma_R f_R(\theta) + \gamma_F f_F(\theta)]$, is only related to the distributions of the latent traits.

IRF $P(\theta)$

The most often used IRT models are the one-parameter logistic (1PL), 2PL, and 3PL models. Consider the IRF for the 3PL model:

$$P_{3PL}(\theta) = c + (1 - c) \frac{1}{1 + \exp(-1.7a(\theta - b))},$$

where b is the difficulty parameter, a is the discrimination parameter, and c is the guessing parameter. The 2PL model is a special case of the 3PL model with $c = 0$, and the Rasch model is a special case of 2PL model with $a = 1$. Figure 1 shows how the product $P(\theta)Q(\theta)$ changes at different values of a , b , and c . In Panels (a) and (c), the item follows 2PL, and in Panels (b) and (d), the item follows the 3PL model.

If the studied item follows the 2PL model, the product $P(\theta)Q(\theta)$ is a bell-shaped function symmetric about $\theta_{1/2}$, where $\theta_{1/2}$ is the value of θ where $P(\theta)$ equals 1/2 (Figure 1, Panels (a) and (c)). It is well known that the 2PL IRF is also the cumulative distribution function (CDF) of a logistic distribution with location parameter b , and scale parameter $1/1.7a$. The derivative of the 2PL IRF is $P'_{2PL}(\theta) = 1.7aP_{2PL}(\theta)Q_{2PL}(\theta)$. Therefore, the product $P(\theta)Q(\theta)$ for the 2PL model can be seen as a PDF of the logistic distribution multiplied by $(1/1.7 \times a)$, and thus has a bell-shaped curve.

However, by including a non-zero guessing parameter c in the 3PL model, the product $P(\theta)Q(\theta)$ is no longer symmetric (Figure 1, Panels (b) and (d)). As $P_{3PL}(\theta) = c + (1 - c)P_{2PL}(\theta)$, it can be derived that $P_{3PL}(\theta)Q_{3PL}(\theta) = c(1 - c)Q_{2PL}(\theta) + (1 - c)^2P_{2PL}(\theta)Q_{2PL}(\theta)$. Therefore, on Panels (b) and (d) of Figure 1, as $\theta \rightarrow -\infty$, the function tends to $c(1 - c)$ (in the figure, $0.2 \times 0.8 = 0.16$), while as $\theta \rightarrow \infty$, the function tends to 0.

Figure 1. $P(\theta)Q(\theta)$ for IRT model at different values of a , b , and c .

Note. IRT = item response theory.

Panels (a) and (b) of Figure 1 show that the difficulty parameter b only affects the location of the curve and does not affect the shape of the curve. Panels (c) and (d) of Figure 1 show that by increasing the discrimination parameter a , the shape of the curve gets narrower.

Latent Trait Distributions

Now consider the term

$$H(\theta) = \frac{f_R(\theta)f_F(\theta)\gamma_R\gamma_F}{\gamma_R f_R(\theta) + \gamma_F f_F(\theta)},$$

which has three components: the proportion of the population composed of the focal group γ_F (reference group, $\gamma_R = 1 - \gamma_F$), the reference group latent trait distribution $f_R(\theta)$, and the focal group latent trait distribution $f_F(\theta)$. Assuming that for the reference group $f_R(\theta) = N(0, 1)$, consider three situations for the focal group $f_F(\theta)$: (a) same distribution as reference group, $p(\theta) = \gamma_R N(0, 1) + \gamma_F N(0, 1) = N(0, 1)$; (b) same variance, different mean, $p(\theta) = \gamma_R N(0, 1) + \gamma_F N(d_0, 1)$; and (c) same mean, different variance, $p(\theta) = \gamma_R N(0, 1) + \gamma_F N(0, s^2)$.

Figure 2 shows the shape of the function $H(\theta)$ under the different situations. When the reference and focal groups have the same population distribution for the latent trait, $f_R(\theta) = f_F(\theta) = p(\theta)$, then $H(\theta) = \gamma_R \gamma_F p(\theta)$. Figure 2, Panel (a) shows the curve $H(\theta)$ when $\gamma_R = \gamma_F = 0.5$, and $p(\theta)$ is the PDF of $N(0, 1)$. Figure 2, Panel (b) shows the curves when the focal group proportion γ_F takes a range of reasonable values, namely, 0.5, 0.3, 0.2, and 0.1. As the proportions of focal and reference groups get more imbalanced, $H(\theta)$ gets smaller, leading to a larger standard error and smaller power. Figure 2, Panel (c) shows the $H(\theta)$ curve when

Figure 2. The shape of $H(\theta)$ under different situations.

the reference and the focal groups have different means but the same variance. Specifically, the reference has a $N(0, 1)$ distribution, and the focal group has a $N(d_\theta, 1)$ distribution, where the curves for $d_\theta = 0, 1, 2, 3$ are plotted. As d_θ increases, the $H(\theta)$ curve has a smaller area under curve, which means the integral tends to get a smaller value, leading to larger standard error and smaller power. Figure 2, Panel (d) shows the curves where the reference and the focal have the same mean but different variances. In this situation, the reference group has a $N(0,1)$ distribution, whereas the focal group has a $N(0, s^2)$ distribution, where curves for $s = 0.5, 1, 2, 3$ are plotted. As s increases, the curve gets flatter. The area under the curve obtains its highest value when the focal group has the same SD as the reference ($s = 1$).

Combining Item Characteristic Curve and Latent Trait Distribution

For the integral $\int P_R(\theta)Q_R(\theta)H(\theta)d\theta$, combining the effects of the item characteristic curve and the test-taker latent trait distribution, one would expect higher power if $P_R(\theta)Q_R(\theta)$ is well matched with the latent trait distribution, as characterized by $H(\theta)$, in terms of the location and the shape of the curves. That means it will be easier to detect DIF for an item with difficulty similar to the mean ability of the population than for those items that are too easy or too difficult.

Table 2. A Summary Table of Factors Influencing the Power of MH Test.

Factor	Change	$ \alpha_{MH} $	SE of $\log \alpha_{MH}$	Power
DIF effect: $b_F - b_R$	$b_F - b_R$ increases by $f_{\Delta b}$ times	Increases by $f_{\Delta b}$ times	Not much change	Increases
Difficulty b_R	$ b_R - M(\theta) $ increases	No change for Rasch and 2PL	Increases	Decreases
Discrimination a	a_R increases by f_a times	Increases by f_a times	Decreases	Increases
Guessing c	c increases	Decreases	Decreases	Decreases
Proportion of focal group in the sample γ_F	$ \gamma_F - 0.5 $ increases from 0 to 0.5	No change	Increases	Decreases
Difference in population mean $d_\theta = \mu_F - \mu_R$	$ d_\theta $ increases	No change for Rasch and 2PL	Increases	Decreases

Note. MH = Mantel-Haenszel; DIF = differential item function; 2PL = two-parameter logistic.

Table 2 summarizes the factors that influence the power of the MH test as discussed so far. These factors were also previously studied and reported by various authors through simulation studies. Here, they were identified from the theoretical power formula. These factors may influence the power in complicated ways. Therefore, the power formulas are important tools to quantify how the power is influenced by these factors. It should be pointed out that the power formulas are very general and their applicability should not be considered as only restricted to the situations already discussed. There is no restriction that the IRF has to be 2PL or 3PL IRT models; actually, it can be of any form, for example, a normal ogive model, even a nonparametric IRT model. Also, the distribution of the latent trait may take forms other than the normal distribution.

Sample Size Formula

Based on the power formula, the formula for sample size can be readily solved. Generally, for a one-sided test, $H_0 : \psi = \psi_0$ versus $H_1 : \psi > \psi_0$, the power formula has the general form $\beta = \Phi(\mu - z_\alpha)$, where $\mu = \sqrt{n}\psi/\tau(\psi_0)$. Solving for μ , $\mu = z_\alpha + z_{1-\beta}$. Note that here α is the Type I error, and β is the power, so $1 - \beta$ is the Type II error. And z_α and $z_{1-\beta}$ correspond to the Z scores in the standard normal distribution with right tail areas α and $1 - \beta$, respectively; namely, $z_\alpha = \Phi^{-1}(1 - \alpha)$ and $z_{1-\beta} = \Phi^{-1}(\beta)$. For example, for a test with significance level $\alpha = .05$, and desired power $\beta = .80$, the corresponding $z_{0.05} = 1.64$ and $z_{0.20} = 0.84$, and $(z_{0.05} + z_{0.20})^2 = 6.15$. Solving for the sample size n ,

$$n = \left[\frac{\tau(\psi_0)}{\psi} (z_\alpha + z_{1-\beta}) \right]^2.$$

By substituting in $\psi = \psi_{MH} = \log \alpha_{MH}$, and the formula for $\tau(\psi_0)$, the formula for the sample size for the one-sided MH test can be obtained as

$$n = \frac{(z_\alpha + z_{1-\beta})^2}{\psi_{MH}^2 \left[\int P_R(\theta) Q_R(\theta) \frac{f_R(\theta) f_F(\theta) \gamma_R \gamma_F}{\gamma_R f_R(\theta) + \gamma_F f_F(\theta)} d\theta \right]}. \quad (7)$$

For the two-sided MH test, $H_0 : \psi = \psi_0$ versus $H_1 : \psi \neq \psi_0$, there is no closed form solution for the sample size. By noting that the power is a monotonic increasing function of the sample size, numeric methods, such as a binary search algorithm, can be used to find the minimal sample size to achieve a desired level of power, and are straightforward to implement.

As an example, for the two-sided MH test with a Type I error rate of 0.05, Table 3 shows the results of minimum sample size needed to achieve 0.80 power when the studied item follows the Rasch model or a model modified from the Rasch model by changing one of the factors. In the default setup (the bolded row in each section of the table), the studied item follows the Rasch model: the reference difficulty b_R is 0, the latent trait distributions for both the focal and reference groups are $N(0,1)$, and the group sizes are balanced ($N_R = N_F = N/2$). Consider two values for $\Delta b = b_F - b_R$: a medium effect value of $\Delta b = 0.426$, and a large effect value of $\Delta b = 0.638$. Under the default setup, if the focal group difficulty is 0.426 (medium effect), the required sample size is 838 ($N_R = N_F = 419$). If the focal group difficulty is 0.638 (large effect), then the required sample size is 374 ($N_R = N_F = 187$). The table shows sample sizes in the cases when one of the following factors changes while keeping the rest fixed at default values: the reference difficulty b_R , discrimination a , guessing c , mean ability μ , mean difference in ability $d_\theta = \mu_R - \mu_F$, and SD of the ability σ . In each case, the asymptotic MH odds ratio α_∞ and its

Table 3. Sample Size Needed to Reach Power 0.80 for the Two-Sided MH Test on the Rasch Model or a Model Modified From the Rasch Model.

										Medium effect				Large effect				
										$\Delta b = b_F - b_R = 0.426$				$\Delta b = b_F - b_R = 0.638$				
										$\log \alpha_{\infty}$	$SE_{\log \alpha} \times \sqrt{n}$	n	$N_R = N_F = n/2$	α_{∞}	$\log \alpha_{\infty}$	$SE_{\log \alpha} \times \sqrt{n}$	n	$N_R = N_F = n/2$
Difficulty for the reference										α_{∞}	$SE_{\log \alpha} \times \sqrt{n}$	n	$N_R = N_F = n/2$	α_{∞}	$\log \alpha_{\infty}$	$SE_{\log \alpha} \times \sqrt{n}$	n	$N_R = N_F = n/2$
$b_R = 0$										1.53	0.426	838	419	1.89	0.638	4.40	374	187
1										1.53	0.426	974	487	1.89	0.638	4.74	434	217
2										1.53	0.426	1,496	748	1.89	0.638	5.88	668	334
3										1.53	0.426	2,892	1,446	1.89	0.638	8.18	1,290	645
Discrimination																		
$a = 0.5$										1.24	0.213	2,932	1,466	1.38	0.319	4.12	1,308	654
1										1.53	0.426	838	419	1.89	0.638	4.40	374	187
2										2.34	0.852	286	143	3.58	1.276	5.14	128	64
3										3.59	1.278	168	84	6.78	1.914	5.90	76	38
Guessing																		
$c = 0$										1.53	0.426	838	419	1.89	0.638	4.40	374	187
0.15										1.41	0.341	1,268	634	1.65	0.502	4.33	586	293
0.2										1.38	0.320	1,446	723	1.60	0.469	4.34	672	336
0.25										1.35	0.301	1,650	825	1.55	0.440	4.36	772	386
Mean ability ($\mu = \mu_R = \mu_F$)																		
0										1.53	0.426	838	419	1.89	0.638	4.40	374	187
1										1.53	0.426	974	487	1.89	0.638	4.74	434	217
2										1.53	0.426	1,496	748	1.89	0.638	5.88	668	334
3										1.53	0.426	2,892	1,446	1.89	0.638	8.18	1,290	645
Mean ability difference ($d_0 = \mu_R - \mu_F$)																		
0										1.53	0.426	838	419	1.89	0.638	4.40	374	187
1										1.53	0.426	1,066	533	1.89	0.638	4.96	476	238
2										1.53	0.426	2,076	1,038	1.89	0.638	6.93	926	463
3										1.53	0.426	5,842	2,921	1.89	0.638	11.62	2,604	1,302
SD of ability ($\sigma = \sigma_R = \sigma_F$)																		
0.5										1.53	0.426	734	367	1.89	0.638	4.12	328	164
1										1.53	0.426	838	419	1.89	0.638	4.40	374	187
2										1.53	0.426	1,144	572	1.89	0.638	5.14	510	255
3										1.53	0.426	1,508	754	1.89	0.638	5.90	672	336

Note. The default values of parameters are $b_R = 0, a = 1, \mu_R = 0, d_0 = 0, \sigma = 1$. The results for the default setup are bolded. In each section, one parameter's value is changed, while the rest stay at the default values.

logarithm $\log \alpha_\infty$, the standard error of $\log \alpha_\infty$ multiplied by $\sqrt{n}$ (which is $\tau(\psi_0)$ in the power formula) are also reported.

Simulation Studies

In this section, the purpose of the simulation studies is to confirm that the power formula is correct, in the sense that the theoretical power is close to the power obtained by Monte Carlo simulation. Here, the theoretical values are compared with the results from two previously published simulation studies. The first was from Paek and Guo (2011), and the second was Uttaro and Millsap (1994). For the Paek and Guo study, their simulation results were used directly, whereas for the Uttaro and Millsap study, their simulation setup was used and the simulation study was re-conducted.

Simulation Study 1: Paek and Guo (2011)

Paek and Guo (2011) examined the power of the MH procedure when the focal and reference groups have notably unbalanced sample size. They focused on the situation where the focal group has a fixed small sample size, which may not satisfy the minimum DIF sample size requirement. They showed through simulation studies that the MH procedure still may have enough power to detect DIF in such situations if the sample size of the reference group is large enough. Here, the results could be reexamined by comparing the DIF detection rates reported in their simulation studies with the theoretical power as calculated by Equation 9.

Paek and Guo (2011) conducted two simulation studies, where in Study 1, they studied the impact of the sample size ratio at a fixed total, and in Study 2, the focal group is fixed at a small sample size whereas the reference group sample size increases. The DIF data were generated from the 3PL model with a scaling factor $D = 1.7$. In each test form, there were 40 anchor items and 1 studied DIF item. For the anchor items, the item parameters were randomly selected from the following distributions: $a \sim N(1, 0.3)$, $b \sim N(0, 1)$, and $c \sim \text{Uniform}(0, 0.35)$. For the studied DIF item: $a_F = a_R = 1.086$, $c_F = c_R = 0.203$, $b_F - b_R = 0.35$, and b_R may take values -0.338 or 0.338 . For person ability, if the group ability differences $d_\theta = 0$, then for both reference and focal group $\theta \sim N(0, 1)$; if $d(\theta) = 0.5$, then $\theta_R \sim N(0.25, 1)$ and $\theta_F \sim N(-0.25, 1)$. By varying b_R , d_θ , N_R , and N_F , in Study 1, there were 44 conditions, and in Study 2, there were 100 conditions. In each condition, 4,000 replicated data sets were simulated. And for each replicated data set, three MH DIF tests: MH- χ^2_c , MH- χ^2 , and MH-Z were applied and the decisions of statistical significance were recorded. The power of the tests in each condition was estimated by the rejection rate (number of significant iterations/total number of iterations). The Monte Carlo error for the power estimates is small (<0.008) due to the large number of iterations. Table 4 (online appendix) reproduces the parameters and the rejection rates as reported in Paek and Guo (2011; Tables 1-2 for Study 1, and Tables 3-6 for Study 2). Based on the parameters used to simulate the data sets in each condition, the theoretical power is calculated from Equation 9, and the values are reported in the last column of Table 4.

The observed rejection rates from the simulated data agree very well with the power calculated from the asymptotic power formula in Equation 9. Panel (a) of Figure 3 shows the scatter plot of the rejection rate from the MH-Z statistic versus the rejection rate from the MH- χ^2 (without continuity correction) statistic, where all the points are very close to the 45° line, reflecting the fact that MH-Z and MH- χ^2 have the same power because they are asymptotically equivalent. Panels (b) and (c) of Figure 3 compare the observed rejection rates from the MH-Z statistic and the MH- χ^2 statistic, respectively, to the theoretical power values calculated by Equation 9. It is clear that the observed rejection rates from the simulation data are close to the

Figure 3. Comparisons of the rejection rate from the Peak and Guo (2011) simulation studies to the theoretical power.
Note. MH = Mantel-Haenszel.

theoretical power as all the points are close to the 45° line. Out of the 144 simulation conditions, the differences between the theoretical power and the simulation rejection rate of the $MH-X^2$ test have a mean of -0.004 and a SD of 0.022 ; whereas the differences for the $MH-Z$ test have a mean of 0.012 and a SD of 0.021 . Panel (d) of Figure 3 compares the rejection rates from $MH-X^2$ statistics with continuity correction to the theoretical power. With the continuity correction, the MH test is conservative in rejecting the null hypothesis and its actual Type I error is lower than the nominal value 0.05 (Paek, 2010). The scatter plot shows that all the observed rejection rates for the continuity-corrected $MH-X^2$ statistic are slightly lower than the theoretical power values. The differences between the theoretical power and the rejection rate of the $MH-X^2_c$ test have a mean of 0.043 and a SD of 0.023 .

Figure 4. Comparison of rejection rates and the theoretical power in re-conducting the Uttaro and Millsap (1994) simulation studies.
 Note. MH = Mantel-Haenszel.

Simulation Study 2: Uttaro and Millsap (1994)

Uttaro and Millsap (1994) conducted a comprehensive simulation study of a variety of factors and their interactions on the detection of DIF by the MH procedure. They designed 20-item and 40-item tests with 36 conditions for the studied item by varying the following four factors: item parameters a , b , c for the focal group in the IRT 3PL model, and the group difference in mean θ level: d_θ . Specifically, they used a $3 \times 3 \times 2 \times 2$ factorial design with (a) three levels of a : 0.5, 1.0, 1.5; (b) three levels of b : 0.0, 0.3, 0.5; (c) two levels of c : 0.0, 0.2; and (d) two levels of d_θ : 0.0, -1.0 . The item parameters for the studied item in the reference group were $a = 1.0$, $b = 0.0$, and $c = 0.2$. The numbers of examinees in the focal and the reference groups were equal: $N_R = N_F = 500$. The latent trait values of the examinees were drawn from $\theta_R \sim N(0, 1)$ and $\theta_F \sim N(d_\theta, 1)$.

In the original article, the number of replicates for each condition was 200, and only the $MH-\chi^2$ test with continuity correction was considered. Instead of directly using the results reported in Uttaro and Millsap (1994), this study re-conducted the simulation studies under the same design, and increased the number of replicates to 2000, and also obtained the results for the $MH-\chi^2$ test without continuity correction and the MH-Z test.

Figure 4 show the scatter plots of rejection rates of the MH tests versus the theoretical power. In the scatter plots, the points are close to the 45° line, indicating the theoretical power values are close to the rejection rates. The differences between the theoretical power value and the rejection rate for 40-item test have a root mean square difference (RMSD) of 0.032 for $MH-\chi^2$ and MH-Z

and an RMSD of 0.041 for MH- χ_c^2 . For the case of 20 items, the corresponding RMSD values are 0.044 (MH-Z), 0.045 (MH- χ_c^2), and 0.052 (MH- χ^2). Based on RMSD, the theoretical power is closer to the 40-item rejection rate than to the 20-item rejection rate. This is expected, because the formula is asymptotic and thus assumes a very long test with a large number of items. However, the increase in RMSD from 40-item (moderate) to 20-item (short) test is quite small from a practical perspective, which indicates that the power formula can still be used for short tests.

Conclusion

In this article, a general formula for calculating the power of the MH test is derived. The formula is asymptotic in the sense that it assumes the number of examinees is large, and the number of the items is large so that by matching the total scores it is very close to the ideal situation that the underlying latent traits are matched. A comparison of the theoretical values of the power and the observed rejection rate in the simulation studies have shown that the formula provides a close approximation to the true power and thus can be used in practical applications. In particular, for example, it can be used in the design of DIF analyses in regard to determining the level power for detecting DIF effect of certain size under current sample size or in determining the sample size needed in each group for detecting a given DIF effect.

The power formula also helps understanding the factors that influence the power of the MH test, as seen by the discussion in this article. There have been many previous simulation studies on the factors that influence the power of the MH test, and the power formula explains precisely why these factors influence the power.

Practitioners should bear in mind that the use of the power formula involves certain limitations. The formula is derived under the assumptions that matching on the total score is equivalent to matching the latent trait. This assumption may not hold if the test is short and the study item is not included as a part of the matching score, and the distributions of the latent trait in the reference and focal group differ, one could see inflated Type I error if the sample size is large (DeMars, 2010; Roussos & Stout, 1996). In such situations, the power is not meaningful because the Type I error is not controlled and, thus, the test itself is not valid.

An R statistical package for conducting the power and sample size calculations for the MH DIF test, using the methods presented in the article, and the package is available upon request.

Acknowledgments

The author thanks Dr. Louis Roussos for his suggestions. She also thanks the editor, the associate editor, and anonymous reviewers for their constructive comments and feedback on the manuscript.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by a Boston College Research Expense grant that was awarded to the author by her employer, Boston College.

Supplemental Material

The online appendix is available at <http://apm.sagepub.com/supplemental>

References

- Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York, NY: John Wiley.
- Birch, M. W. (1964). The detection of partial association, I: The 2×2 case. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26, 313-324.
- Cheng, Y., Chen, P., Qian, J., & Chang, H.-H. (2013). Equated pooled booklet method in DIF testing. *Applied Psychological Measurement*, 37, 276-288.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- DeMars, C. E. (2010). Type I error inflation for detecting DIF in the presence of impact. *Educational and Psychological Measurement*, 70, 961-972.
- Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and practice* (pp. 137-166). Hillsdale, NJ: Lawrence Erlbaum.
- Finch, W. H. (2012). The MIMIC model as a tool for differential bundle functioning detection. *Applied Psychological Measurement*, 36, 40-59.
- 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: Lawrence Erlbaum.
- Holland, P. W., & Wainer, H. E. (1993). *Differential item functioning: Theory and practice*. Hillsdale, NJ: Lawrence Erlbaum.
- Lehmann, E. L. (1999). *Elements of large-sample theory*. New York, NY: Springer.
- Li, Z. (2014a). Power and sample size calculations for logistic regression tests for differential item functioning. *Journal of Educational Measurement*, 51, 441-462.
- Li, Z. (2014b). A power formula for the SIBTEST procedure for differential item functioning. *Applied Psychological Measurement*, 38, 311-328.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719-748.
- Millisap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, 17, 297-334.
- Paek, I. (2010). Conservativeness in rejection of the null hypothesis when using the continuity correction in the MH chi-square test in DIF applications. *Applied Psychological Measurement*, 34, 539-548.
- Paek, I., & Guo, H. (2011). Accuracy of DIF estimates and power in unbalanced designs using the Mantel-Haenszel DIF detection procedure. *Applied Psychological Measurement*, 35, 518-535.
- Phillips, A., & Holland, P. W. (1987). Estimators of the variance of the Mantel-Haenszel log-odds-ratio estimate. *Biometrics*, 43, 425-431.
- R Core Team. (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available from <http://www.r-project.org>
- 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, 293-322.
- Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement*, 33, 215-230.
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27, 361-370.
- Uttaro, T., & Millisap, R. E. (1994). Factors influencing the Mantel-Haenszel procedure in the detection of differential item functioning. *Applied Psychological Measurement*, 18, 15-25.
- Wang, W., Tay, L., & Drasgow, F. (2013). Detecting differential item functioning of polytomous items for an ideal point response process. *Applied Psychological Measurement*, 37, 316-335.
- Zwick, R. J. (2012). *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement* (ETS Research Report No. RR-12-08). Princeton, NJ: ETS.