The Effect of Upper and Lower Asymptotes of IRT Models on Computerized Adaptive Testing

Applied Psychological Measurement
2015, Vol. 39(7) 551–565
© The Author(s) 2015
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621615585850
apm.sagepub.com

Ying Cheng¹ and Cheng Liu¹

Abstract

In this article, the effect of the upper and lower asymptotes in item response theory models on computerized adaptive testing is shown analytically. This is done by deriving the step size between adjacent latent trait estimates under the four-parameter logistic model (4PLM) and two models it subsumes, the usual three-parameter logistic model (3PLM) and the 3PLM with upper asymptote (3PLMU). The authors show analytically that the large effect of the discrimination parameter on the step size holds true for the 4PLM and the two models it subsumes under both the maximum information method and the *b*-matching method for item selection. Furthermore, the lower asymptote helps reduce the positive bias of ability estimates associated with early guessing, and the upper asymptote helps reduce the negative bias induced by early slipping. Relative step size between modeling versus not modeling the upper or lower asymptote under the maximum Fisher information method (MI) and the *b*-matching method is also derived. It is also shown analytically why the gain from early guessing is smaller than the loss from early slipping when the lower asymptote is modeled, and vice versa when the upper asymptote is modeled. The benefit to loss ratio is quantified under both the MI and the *b*-matching method. Implications of the analytical results are discussed.

Keywords

computerized adaptive testing, early slipping, early guessing, upper asymptote, lower asymptote, (relative) step size

When the items are scored dichotomously, such as right/wrong, dichotomous item response theory (IRT) models are often used to model the relationship between the underlying latent trait θ and a respondent's score. According to the widely used three-parameter logistic model (3PLM; Birnbaum, 1968), the probability for an examinee with latent trait level θ to receive score $u_j = 1$ on the j th item ($u_j = 1$ for a correct response and 0 for an incorrect one) is

$$P_j(\theta) = P(u_j = 1 | \theta, \gamma_j) = c_j + (1 - c_j) \frac{e^{a_j(\theta - b_j)}}{1 + e^{a_j(\theta - b_j)}}, \quad (1)$$

¹University of Notre Dame, IN, USA

Corresponding Author:

Ying Cheng, Department of Psychology, University of Notre Dame, 118 Haggard Hall, Notre Dame, IN 46556, USA.
Email: ycheng4@nd.edu

where the parameters $\gamma_j = (a_j, b_j, c_j)'$ are item discrimination, difficulty, and pseudo-guessing parameters, respectively. The probability of giving an incorrect answer, $Q_j(\theta)$, equals $1 - P_j(\theta)$. The nomenclature was motivated by educational testing. An item with larger b_j leads to lower $P_j(\theta)$, given θ and therefore is more "difficult." In educational testing, examinees may try to "guess" at the right answer. Hence, the parameter c_j is called the "pseudo-guessing" parameter or sometimes the "guessing" parameter for short. If the lower asymptotes are set to 0, the 3PLM reduces to the two-parameter logistic model (2PLM). If the discrimination parameters are further constrained to be equal across items, the model reduces to the one-parameter logistic model (1PLM). When all discrimination parameters are constrained to be 1, the model is referred to as the Rasch model.

As early as 1981, Barton and Lord discussed adding an upper asymptote, d , to the 3PLM to accommodate unfortunate slips of highly capable examinees, similar to the lower asymptote accounting for lucky guessing from low-ability examinees. The resulting model is

$$P_j(\theta) = c_j + (d - c_j) \frac{e^{a_j(\theta - b_j)}}{1 + e^{a_j(\theta - b_j)}}. \quad (2)$$

A more general model, the four-parameter logistic model (4PLM), further allows the upper asymptotes to differ across items

$$P_j(\theta) = P(u_j = 1 | \theta, \gamma_j) = c_j + (d_j - c_j) \frac{e^{a_j(\theta - b_j)}}{1 + e^{a_j(\theta - b_j)}}. \quad (3)$$

For item j , the item parameter vector is $\gamma_j = (a_j, b_j, c_j, d_j)'$. Model 2 is a constrained version of the 4PLM in Equation 3 when all $d_j \equiv d$. A normal-ogive version of the four-parameter model was first proposed by McDonald (1967). It includes a lower and an upper asymptote for each item.

Chang and Ying (2008) provided important theoretical results on how the step size between adjacent latent trait estimates is influenced by n , the number of items administered, and the discrimination parameter a . Their analytical results show that when highly discriminating items are used early in the test (i.e., when n is still small), the one-step adjustment or step size can be very large. Therefore, they projected that

if the examinee misses a number of initial items and the test length is short to moderate, then he or she may not be able to regain a score comparable to the true θ , even though he or she responds well to the rest of the items (p. 444).

Their simulation study further supported this assertion and showed that the a -stratify method, by using low- a items early in the test, offers a more robust algorithm for item selection in computerized adaptive testing (CAT).

The objective of this article is to extend the analytical results in Chang and Ying (2008). The extension happens in three ways. First, Chang and Ying focused on the role of the discrimination parameter in step size. In this article, the role of the upper and lower asymptotes is examined. The authors show analytically that these asymptotes can reduce the step size between adjacent interim ability estimates in CAT. Previous simulation studies (e.g., Rulison & Loken, 2009) alluded to the effects of asymptote parameters but analytical results are warranted. Second, the results of the present study are no longer limited to the maximum information method for item selection. General results, as well as results specifically under the maximum information method and the b -matching method, are obtained. Third, relative step size between

modeling versus not modeling the asymptotes is derived. The authors also explain analytically the asymmetry, that is, larger gain from early guessing than loss from early slipping when the upper asymptote is included in the model, and vice versa when the lower asymptote is included in the model.

The rest of the article is organized as follows. First, the authors discuss technical aspects of CAT under the 4PLM, the model with both lower and upper asymptotes. The discussion focuses primarily on ability update and item selection, which are instrumental to the step size. Then, the authors derive the step size between successive ability updates under the 4PLM and the two models it subsumes, the model with the lower asymptote only (3PLM), and the model with upper asymptote only (3PLMU). The step size and relative step size are derived under both the maximum information method and the b -matching method. Implications of the analytical results are discussed at the end.

CAT Based on the 4PLM

Barton and Lord (1981) fit Model 2 to data of large-scale aptitude tests, and the conclusion was that the change in ability estimates was negligible. Since then, the discussion of adding an upper asymptote to the IRT model was rare, until very recently when there seems to be revived interest in this topic. There are three main reasons for such rekindled interest. First, the upper asymptote was found valuable in fitting psychological testing data (Loken & Rulison, 2010; Reise & Waller, 2003; Waller & Reise, 2009). Second, failing to account for unfortunate slips early in a test due to carelessness or warm-up effects may lead to negative consequences in CAT (Rulison & Loken, 2009; Yen, Ho, Laio, Chen, & Kuo, 2012). Last but not the least, in Barton and Lord (1981), the upper asymptote was fixed across items. Recent advances in computational statistics have made it straightforward to estimate an upper asymptote for each item (Loken & Rulison, 2010), that is, Model 3.

Liao, Ho, Yen, and Cheng (2012) and Yen et al. (2012) discussed computerized adaptive testing that is built upon the Barton and Lord (1981) model, where the upper asymptote is fixed across items. Below, it has been shown that traditional item selection and ability estimation methods can be easily modified to be used in CAT built on the more general 4PLM in Model 3.

There are four important components to a CAT (Cheng & Keng, 2009): (a) the start rule, that is, what item or items to give to examinees to start the test with. Oftentimes, the first items are chosen randomly in the bank or in a certain difficulty range of the bank; (b) the ability estimation method that updates $\hat{\theta}$, the provisional estimate of θ every time one more item is answered; (c) the item selection method that determines which item in the bank to administer next based on the interim $\hat{\theta}$; and (d) the termination rule that determines when to stop the test. The choice of the first items is usually not adaptive so it is not the focus here. The authors also focus on fixed-length CAT, so the test terminates when the predetermined test length is reached. Below ability estimation and item selection under the 4PLM are discussed.

Ability Estimation

Assuming local independence, the likelihood function given a person's response vector $\mathbf{u} = (u_1, u_2, \dots, u_m)'$ of m items is

$$L(\theta | \mathbf{u}, \gamma) = \prod_{j=1}^m P_j(\theta)^{u_j} Q_j(\theta)^{(1-u_j)}, \quad (4)$$

where $\boldsymbol{\gamma}$ represents the collection of all $\boldsymbol{\gamma}_j$, $j = 1, 2, \dots, m$, and $P_j(\theta)$ is defined in Equation 3 and are assumed known. The maximum likelihood estimate of θ , denoted as $\hat{\theta}_{ML}$, maximizes the likelihood or equivalently, the log likelihood:

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}} \{ \log L(\theta | \mathbf{u}, \boldsymbol{\gamma}) \} = \underset{\theta}{\operatorname{argmax}} \left\{ \sum_{j=1}^m [u_j \log(P_j(\theta)) + (1 - u_j) \log(Q_j(\theta))] \right\}.$$

Obtaining the $\hat{\theta}_{ML}$ therefore involves solving the estimation equation

$$\sum_{j=1}^m \left[u_j \frac{\partial \log(P_j(\hat{\theta}_{ML}))}{\partial \theta} + (1 - u_j) \frac{\partial \log(Q_j(\hat{\theta}_{ML}))}{\partial \theta} \right] = 0.$$

The Newton–Raphson iterative procedure can be used to solve the above equation, which involves the first- and second-order derivative of the log likelihood (Embretson & Reise, 2000). The first- and second-order derivatives under the 4PLM are given in Online Appendix A.

A popular alternative is the Expected a Posterior (EAP) estimate, $\hat{\theta}_{EAP}$. Assuming a prior distribution for the latent trait, $\lambda(\theta)$,

$$\hat{\theta}_{EAP} = \frac{\int \theta L(\theta | \mathbf{u}, \boldsymbol{\gamma}) \lambda(\theta) d\theta}{\int L(\theta | \mathbf{u}, \boldsymbol{\gamma}) \lambda(\theta) d\theta}.$$

Computationally, the integration is transformed into a summation of weighted likelihoods at quadrature points:

$$\hat{\theta}_{EAP} = \frac{\sum_{k=1}^{n_q} s_k L(s_k | \mathbf{u}, \boldsymbol{\gamma}) w(s_k)}{\sum_{k=1}^{n_q} L(s_k | \mathbf{u}, \boldsymbol{\gamma}) w(s_k)},$$

where s_k is one of the quadrature points, n_q is the number of quadrature points, $w(s_k)$ is the weight associated with the quadrature point under the prior distribution, and $L(s_k | \mathbf{u}, \boldsymbol{\gamma})$ is the likelihood function evaluated at the quadrature point (see Equation 4). By having a large number of numerical quadrature points, the approximation can be very accurate.

Item Selection

Once the ability estimate is updated, the CAT program selects among eligible items in the bank the most suitable one as the next item to administer. The most widely known item selection method in CAT is the maximum Fisher information method (MI; Lord, 1980). Eligible items in the bank are rank ordered according to their item information at the most recent $\hat{\theta}$, and the item with the largest item information gets chosen. Magis (2013) gives the information of item j under the 4PLM in the bank at $\hat{\theta}$:

$$I_j(\hat{\theta}) = \frac{a_j^2 [P_j(\hat{\theta}) - c_j]^2 [d_j - P_j(\hat{\theta})]^2}{(d_j - c_j)^2 P_j(\hat{\theta}) Q_j(\hat{\theta})}. \quad (5)$$

It is straightforward to compute the item information at $\hat{\theta}$ for all eligible items in the bank following Equation 5, and find the one with the largest information. When $d_j \equiv 1$, the item

information in Equation 5 reduces to item information for the 3PLM. When $d_j \equiv 1$ and $c_j \equiv 0$, the item information in Equation 5 further reduces to item information for the 2PLM.

The MI method under the 2PLM or the 3PLM has been criticized for favoring highly discriminating items, because the information function includes the term a_j^2 . This results in unbalanced item exposure: Some items are used frequently, which pose threats to test security as they become known to prospective students, and some items are seldom used or never used, which is a waste of resources. Because the information function in Equation 5 still includes a_j^2 , the authors expect the MI method under the 4PLM to favor highly discriminating items as well.

Another criticism of the MI method is related to early slipping in CAT. Chang and Ying (2008) showed mathematically that the adjustment step size between $\hat{\theta}^{(t)}$ (the θ estimate after t items have been administered) and $\hat{\theta}^{(t+1)}$ (the θ estimate after $(t+1)$ items have been administered) is proportional to the discrimination parameter of the item chosen as the $(t+1)$ th item for the examinee. As a result, when early slips happen, the ability estimate moves to the low end of the θ continuum rapidly with large steps. Chang and Ying (2008) observed that highly discriminating items are used early in a CAT under the MI method. Therefore, the ability estimate is negatively biased when early slips happen even if the examinee answers the later items correctly, as the early bias may be too large to recover from. In fact, many reported that answers to the first several items are key to the success or failure on a test (Kamat, 2013; Koprince, 2012; Sackmann, 2010). This motivated the “ascending a ” principle, meaning to use items with small discrimination parameters early and items with a parameters later.

As kindly pointed out by a reviewer, given the well-known heavy reliance on highly discriminating items of the MI method, other item selection algorithms should be considered. Here, the case where the $(t+1)$ th item is chosen through b -matching, that is, the match-ability-with-difficulty method (Hulin, Drasgow, & Parsons, 1983), is considered. Note that such strategy ignores the discrimination parameter in item selection and therefore does not favor highly discriminating items.

Rulison and Loken (2009) argued that the 4PLM helps reduce the negative bias caused by early slipping in CAT, and demonstrated it using simulation studies given the *constrained* 4PLM of Barton and Lord (1981) under the MI method. The goal of this study is to show the effects of the asymptotes *analytically* under both the MI and the b -matching method.

Theoretical Results on Step Size Between Successive Interim Ability Estimates

Chang and Ying (2008) mathematically derived the update step size between $\hat{\theta}^{(t)}$ and $\hat{\theta}^{(t+1)}$ under the 1PLM, 2PLM, and 3PLM when the MI method is used. The derivation suggests that in the early stage of a test, the discrimination parameter of the $(t+1)$ th item chosen for the examinee (denoted by $a^{(t+1)}$) has a large influence on the step size. Examinees with early slips would be pushed to such low extreme of the θ scale that they could not recover even with successes on later items. One would expect that similarly, early lucky guesses would push examinees to the high end of the θ scale. Rulison and Loken’s (2009) simulation study confirmed that under the 3PLM, the early slips of high-ability examinees did result in negatively biased ability estimates. However, the lucky guessers did not benefit much in their final ability estimates. They suggested that such asymmetry occurred because 3PLM explicitly accounted for lucky guessing but not slipping.

According to Chang and Ying (2008), the step size between $\hat{\theta}^{(t)}$ and $\hat{\theta}^{(t+1)}$ under the 3PLM can be approximated by

$$\frac{a^{(t+1)}}{I^{(t+1)}(\theta^*)} g\left(c^{(t+1)}\right) \left[u^{(t+1)} - P^{(t+1)}\left(\hat{\theta}^{(t)}\right) \right], \quad (6)$$

where $a^{(t+1)}$ is the discrimination parameter of the $(t+1)$ th item. θ^* is a point between $\hat{\theta}^{(t)}$ and $\hat{\theta}^{(t+1)}$, and $I^{(t+1)}(\theta^*)$ is the test information at the $(t+1)$ th item evaluated at θ^* , that is, the summation of the item information evaluated at θ^* of all the $(t+1)$ items administered. And $P^{(t+1)}(\hat{\theta}^{(t)})$ is the probability of getting the $(t+1)$ th item correctly given the ability estimate $\hat{\theta}^{(t)}$. In addition, $g(c) = (1 + \sqrt{1+8c}) / (2c + 1 + \sqrt{1+8c})$. If the answer to the $(t+1)$ th item is correct, that is, $u^{(t+1)} = 1$, $\left[u^{(t+1)} - P^{(t+1)}(\hat{\theta}^{(t)}) \right]$ must be non-negative, and $\hat{\theta}^{(t+1)}$ must be larger than $\hat{\theta}^{(t)}$ by the amount given by Equation 6. Similarly, an incorrect answer will lead to a $\hat{\theta}^{(t+1)}$ smaller than $\hat{\theta}^{(t)}$ by due amount. Based on Equation 6, Chang and Ying (2008) concluded that the step size is largely governed by discrimination parameter early on in the test when the test information $I(\theta^*)$ is small.

Similar to Chang and Ying (2008), it can be shown that the step size under the 4PLM can be approximated by

$$\frac{a^{(t+1)}}{I^{(t+1)}(\theta^*)} h\left(c^{(t+1)}, d^{(t+1)}\right) \left[u^{(t+1)} - P^{(t+1)}\left(\hat{\theta}^{(t)}\right) \right], \quad (7)$$

where $h(c^{(t+1)}, d^{(t+1)})$ is a function of the lower and upper asymptote of the $(t+1)$ th item administered; $h(c, d)$ is expected to fall between 0 and 1, for any $d > c$ (which is a very reasonable condition, otherwise the item response function (IRF) is invalid). In Online Appendix C, it is shown that

$$h(c, d) = \frac{d - c}{c(1 - c)\left(\frac{d-x^*}{x^*-c}\right) + d(1 - d)\left(\frac{x^*-c}{d-x^*}\right) + (d + c - 2cd)},$$

where x^* is given in Equations 13 and 18 of Magis (2013), and $x^* \in (c, d)$.

To show that $h(c, d)$ falls between 0 and 1 for any $d > c$, first, it is shown that $h(c, d) > 0$. The numerator, $d - c > 0$. The denominator has three components, $z_1 = c(1 - c)((d - x^*)/(x^* - c))$, $z_2 = d(1 - d)((x^* - c)/(d - x^*))$, and $z_3 = d + c - 2cd$. For $x^* \in (c, d)$, obviously z_1 and z_2 are both positive. Because both c and d are between 0 and 1,

$$z_3 = d + c - 2cd \geq d^2 + c^2 - 2cd = (d - c)^2 \geq 0.$$

Therefore, the denominator ($z_1 + z_2 + z_3$) > 0 . When the numerator and denominator are both positive, the ratio must be positive. So $h(c, d) > 0$.

Next, it is shown that $h(c, d) < 1$. Because $z_3 = d + c - 2cd = d - c + 2c - 2cd = (d - c) - 2c(1 - d) > (d - c)$, and z_1 and z_2 are clearly positive, the denominator ($z_1 + z_2 + z_3$) must be larger than $(d - c)$. Then it is obvious that the ratio, $h(c, d)$, is smaller than 1. In summary, $h(c, d)$ is between 0 and 1.

It is clear from Equation 7 that the size of the discrimination parameter still has large influence on the early step sizes in CAT under the 4PLM. Meanwhile, the step sizes are moderated by $h(c, d)$, which is always between 0 and 1. This helps explain why the use of 4PLM can improve the ability estimation of examinees with early slips and early guessing (Rulison & Loken, 2009).

Below, the effect of c and d in isolation has been examined. When the model only includes the lower asymptote c , that is, when $d = 1$, the 4PLM reduces to the regular 3PLM. When the

model only includes the upper asymptote d , that is, when $c = 0$, the 4PLM reduces to the following:

$$P_j(\theta) = P(u_j = 1 | \theta, \gamma_j) = d_j \frac{e^{a_j(\theta - b_j)}}{1 + e^{a_j(\theta - b_j)}}. \quad (8)$$

This model is referred to as 3PLMU, that is, the three-parameter model with an upper asymptote.

Effect of the Lower Asymptote on Step Size

To isolate the effect of the lower asymptote, the step size under the 3PLM, which is given in Equation 6, is examined. Chang and Ying (2008) showed that

$$g(c) = \frac{1 + \sqrt{1 + 8c}}{2c + 1 + \sqrt{1 + 8c}}. \quad (9)$$

Let $\sqrt{1 + 8c} = r$. Because $c \in (0, 1)$, it is obvious that $r \in (1, 3)$, and r increases in c . Equation 9 can be rewritten as a function of r :

$$g(r) = \frac{1 + r}{\frac{r^2 - 1}{4} + 1 + r} = \frac{4}{r + 3}. \quad (10)$$

Clearly, $g(r)$ is decreasing in r , and therefore, decreases in c . So as c increases from 0 to 1, $g(c)$ drops from 1 to 2/3. Given, a c value, $g(c)$ is bounded between 2/3 and 1. The fact that $g(c)$ is smaller than 1, when $c > 0$ shows that modeling the c parameter can reduce the step size.

(Relative) step size under the maximum information method under the 3PLM. Assume that t items have been administered and the current ability estimate is $\hat{\theta}^{(t)}$. The next question can be selected through the maximum information method. In that case, the $(t + 1)$ th item should satisfy (if the item pool is rich enough; Birnbaum, 1968):

$$b^{(t+1)} = \hat{\theta}^{(t)} - \frac{1}{a^{(t+1)}} \ln \left(\frac{1 + \sqrt{1 + 8c^{(t+1)}}}{2} \right), \quad (11)$$

and the probability of answering the the $(t + 1)$ th item correctly is expected to be $(1 + \sqrt{1 + 8c^{(t+1)}})/4$ (Lord, 1980).

Following Equation 6, the step size between $\hat{\theta}^{(t+1)}$ and $\hat{\theta}^{(t)}$ under the MI method, when lucky guessing happens on the $(t + 1)$ th item, is

$$\frac{a_{MI}^{(t+1)}}{I^{(t+1)}(\theta^*)} g(c_{MI}^{(t+1)}) \left(\frac{3 - \sqrt{1 + 8c_{MI}^{(t+1)}}}{4} \right), \quad (12)$$

where the subscript MI indicates that the item is selected under the MI method. In general, the items preferred under the MI method have large discrimination and small guessing parameters. In other words, $a_{MI}^{(t+1)}$ tend to be big, and $c_{MI}^{(t+1)}$ tend to be small. Note that $(3 - \sqrt{1 + 8c_{MI}^{(t+1)}})/4$ is a decreasing function in the lower asymptote. So does $g(c_{MI}^{(t+1)})$. Therefore, big discrimination parameter and small lower asymptote tend to enlarge the step size in Equation 12, when lucky guessing happens. Chang and Ying (2008) focused on the effect of the discrimination parameter, but here, it is shown from Equation 12 that when the MI method

is used, its preference of low- c items would lead to a boost in the ability estimate in the presence of lucky guessing as well.

Other things being equal, what is the effect by simply including the lower asymptote on the $(t+1)$ th item? To find out the effect, a is fixed and the effect of c is isolated. Note that all t items that have been administered are the same. The difference only lies in the $(t+1)$ th item, whether the $c^{(t+1)}$ is either a free parameter between 0 and 1 or not modeled (i.e., fixed at 0). Hence, the test information function in Equation 12, which is the summation of the item information of all the $(t+1)$ items, only differs in terms of the item information of the $(t+1)$ th item between modeling or not modeling the lower asymptote on the $(t+1)$ th item. The test information of the first t items should dominate over the item information of the $(t+1)$ th item. Therefore, the ratio of the test information between modeling or not modeling the lower asymptote parameter of the $(t+1)$ th item ≈ 1 . Then, given a , just by including a lower asymptote parameter between 0 and 1 in the model, the step size can be reduced, because obviously $g(c_{MI}^{(t+1)})(3 - \sqrt{1 + 8c_{MI}^{(t+1)}})/4$ monotonically decreases in c , and its maximum of $1/2$ is reached when $c = 0$. In fact, the step size is reduced to approximately $g(c_{MI}^{(t+1)})(3 - \sqrt{1 + 8c_{MI}^{(t+1)}})/2$ times its original size (i.e., the step size when c is fixed to 0 or not modeled). For example, for $c_{MI}^{(t+1)} = .25$, the step size reduces to 53.6% of its original size, when the examinee answers the $(t+1)$ th item right by lucky guessing. In other words, the relative step size = .536. Consider the extreme case, that is, $c_{MI}^{(t+1)} = 1$. The step size becomes 0, meaning there will not be any update of the ability estimate because the item provides no information at all. The solid curve in the left panel of Figure 1 shows how the relative step size changes as a function of c when items are chosen with the MI method. The points (.25, .536) and (1, 0) clearly are on the curve.

(Relative) step size under b-matching under the 3PLM. If the item pool is rich enough, $b^{(t+1)} = \hat{\theta}^{(t)}$, and the probability of answering the $(t+1)$ th item correctly is $(1 + c^{(t+1)})/2$. Then step size in the presence of lucky guessing is

$$\frac{a_{BM}^{(t+1)}}{I^{(t+1)}(\theta^*)} g\left(c_{BM}^{(t+1)}\right) \left(\frac{1 - c_{BM}^{(t+1)}}{2}\right), \quad (13)$$

and the subscript BM indicates that the item is selected under the b -matching method. From Equation 13, it is clear that under the b -matching method, if the CAT is operating under an IRT model that is not 1PLM, then the discrimination parameter still plays an important role in the step size. Fortunately under the b -matching method, there is no heavy reliance on the highly discriminating items early in the test, so the step size is not expected to be very big in the beginning and to decrease as test progresses.

It is probably of more interest to the readers how the lower asymptote affects step size, when lucky guessing happens. Equation 13 shows that having the lower asymptote makes the step size smaller, because $g(c_{BM}^{(t+1)})(1 - c_{BM}^{(t+1)})/2$ is a decreasing function of the lower asymptote, and its maximum of $1/2$ is reached when $c = 0$. The step size can be reduced to approximately $g(c_{BM}^{(t+1)})(1 - c_{BM}^{(t+1)})$ times its original size. Here, original size refers to the step size when the lower asymptote is not modeled (or simply set to be 0), which is $1/2$. When $c_{BM}^{(t+1)} = .25$, the step size is reduced to 63.4% of its original size.

Interestingly, having a lower asymptote parameter helps reduce the step size to a larger degree under the MI method than the b -matching method, as shown by the case when $c = .25$. In fact, for the same item, the ratio of step size under the MI versus the b -matching strategy is $((3 - \sqrt{1 + 8c^{(t+1)}})/4)/((1 - c^{(t+1)})/2) = (3 - \sqrt{1 + 8c^{(t+1)}})/2(1 - c^{(t+1)})$. When $c > 0$, $(3 - \sqrt{1 + 8c^{(t+1)}})/2(1 - c^{(t+1)}) < 1$. So the relative step size is smaller under the MI method. In

Figure 1. Relative step size between modeling versus not modeling the lower (left panel) or the upper (right panel) asymptote.

Note. MI = maximum Fisher information method; BM = *b*-matching method.

other words, having the lower asymptote in the model helps reduce the step size to a larger extent under the MI method than under the *b*-matching method. The dotted curve in the left panel of Figure 1 shows how the relative step size changes with c when the *b*-matching method is used for item selection. It dominates the solid curve (when the MI method is used for item selection) for $\forall c \in (0, 1)$.

Asymmetry: Early guessing versus early slipping. It can also be shown analytically why under the 3PLM and MI method, the benefit of early lucky guesses is not as much as the loss of early slips, as demonstrated by simulations in Rulison and Loken (2009). The rationale is as follows. Early on in the test when the test information $I(\theta^*)$ is small, the stepwise update is largely governed by the product of $a^{(t+1)}g(c^{(t+1)})[u^{(t+1)} - P^{(t+1)}(\hat{\theta}^{(t)})]$. If the answer to the $(t+1)$ th item is correct, the product is $a^{(t+1)}g(c^{(t+1)})[1 - P^{(t+1)}(\hat{\theta}^{(t)})]$. Compared with $c = 0$, a positive c leads to $g(c^{(t+1)})$ smaller than 1, and also a smaller $[1 - P^{(t+1)}(\hat{\theta}^{(t)})]$. So the step size is reduced by $g(c^{(t+1)})$ and further by $[1 - P^{(t+1)}(\hat{\theta}^{(t)})]$ when a correct answer is given. This explains why the effect of lucky guesses is reduced under the 3PLM. If the answer is incorrect, however, the stepwise adjustment is $a \cdot g(c^{(t+1)})[-P^{(t+1)}(\hat{\theta}^{(t)})]$. Disregarding the sign, the step size obtained is $a \cdot g(c^{(t+1)})[P^{(t+1)}(\hat{\theta}^{(t)})]$. Compared with $c = 0$, a positive c still leads to $g(c^{(t+1)})$ smaller than 1, but $[P^{(t+1)}(\hat{\theta}^{(t)})]$ increases. So the step size toward the low end of the ability scale due to an incorrect answer is larger than the step size toward the high end due to a correct answer, given a positive c . This helps explain the asymmetry, that is, under the 3PLM, examinees do not benefit from early lucky guesses as much as they suffer from early slips.

Formally, the step size under the MI method when a correct answer from lucky guess is registered is given in Equation 12. However, from Equation 6, the step size when an incorrect answer is given to the last item due to slipping is,

$$\frac{a_{MI}^{(t+1)}}{I^{(t+1)}(\theta^*)} g\left(c_{MI}^{(t+1)}\right) \left(\frac{1 + \sqrt{1 + 8c_{MI}^{(t+1)}}}{4}\right). \quad (14)$$

Consequently, the ratio of the step size under guessing versus slipping is $(3 - \sqrt{1 + 8c_{MI}^{(t+1)}})/(1 + \sqrt{1 + 8c_{MI}^{(t+1)}})$, which is smaller than 1 for $\forall c_{MI}^{(t+1)} \in (0, 1)$ (proof omitted). This formally shows that the benefit of early lucky guesses is not as much as the loss of early slips given a positive lower asymptote.

If the b -matching method is used for item selection, the step size under lucky guessing is given in Equation 13. The step size under slipping is,

$$\frac{a_{BM}^{(t+1)}}{I^{(t+1)}(\theta^*)} g\left(c_{BM}^{(t+1)}\right) \left(\frac{1 + c_{BM}^{(t+1)}}{2}\right). \quad (15)$$

The ratio is therefore $(1 - c_{BM}^{(t+1)})/(1 + c_{BM}^{(t+1)})$. Again the ratio is smaller than 1 for $\forall c_{BM}^{(t+1)} \in (0, 1)$. This formally shows that the benefit of early lucky guesses is not as much as the loss of early slips given a positive lower asymptote, when the b -matching method is used for item selection as well. In addition, the authors not only show analytically whether and how the asymmetry occurs but also quantify the relative benefit to loss. The ratio is $(3 - \sqrt{1 + 8c_{MI}^{(t+1)}})/(1 + \sqrt{1 + 8c_{MI}^{(t+1)}})$ and $(1 - c_{BM}^{(t+1)})/(1 + c_{BM}^{(t+1)})$ for the MI or the b -matching method, respectively. So far the effect of the lower asymptote in the 3PLM on the step size between adjacent ability estimates has been shown. Next, the authors show that the upper asymptote in the 3PLMU has a complementary impact.

Effect of the Upper Asymptote on Step Size

The model that contains only the upper asymptote is the 3PLMU model, shown in Equation 8. Following Chang and Ying (2008), it can be shown that the step size under the 3PLMU is approximately,

$$\frac{a^{(t+1)}}{I^{(t+1)}(\theta^*)} f\left(d^{(t+1)}\right) \left[u^{(t+1)} - P^{(t+1)}(\hat{\theta}^{(t)})\right], \quad (16)$$

where $f(d) = (3 - \sqrt{9 - 8d})/2d$. For details of the derivation, please see Online Appendix B. Note that $f(d)$ falls between $2/3$ and 1 , and is monotone increasing in d (proof omitted). When d drops from 1 , $f(d)$ drops, and reduces the step size. Therefore, it is expected that an upper asymptote smaller than 1 reduces the negative bias induced by early slipping.

When early slips happen, the step size is $((a^{(t+1)})/(I^{(t+1)}(\theta^*)))f(d^{(t+1)})[P^{(t+1)}(\hat{\theta}^{(t)})]$. Early in the test, $I(\theta^*)$ is still small so the step size is largely governed by the product of $a^{(t+1)}f(d^{(t+1)})[P^{(t+1)}(\hat{\theta}^{(t)})]$. Any d between 0 and 1 will lead to $f(d) < 1$, and a $P(\hat{\theta})$ that is smaller than when $d = 1$. As a result, the step size is smaller than if $d = 1$. Smaller d leads to smaller $f(d^{(t+1)})[P^{(t+1)}(\hat{\theta}^{(t)})]$. This helps explain why the 3PLMU improves ability estimation for examinees with early slips.

(Relative) step size under the MI method under the 3PLMU. Under the 3PLMU, the next item to be selected should satisfy (if the item pool is rich enough)

$$b^{(t+1)} = \hat{\theta}^{(t)} + \frac{1}{a^{(t+1)}} \ln\left(\frac{2}{1 + \sqrt{9 - 8d^{(t+1)}}}\right), \quad (17)$$

and the probability of answering the next item correctly is expected to be $(3 - \sqrt{9 - 8d^{(t+1)}})/4$ (see Online Appendix B for the proof).

Following Equation 16, the step size between $\hat{\theta}^{(t+1)}$ and $\hat{\theta}^{(t)}$ under the MI method, when slipping happens on the $(t + 1)$ th item, is

$$\frac{a_{MI}^{(t+1)}}{I^{(t+1)}(\theta^*)} f(d_{MI}^{(t+1)}) \left(\frac{3 - \sqrt{9 - 8d_{MI}^{(t+1)}}}{4} \right). \quad (18)$$

Note that the step size is always positive as it quantifies the magnitude, not the direction. When slipping happens, $\hat{\theta}^{(t+1)}$ is smaller than $\hat{\theta}^{(t)}$ by the magnitude quantified in Equation 18. In general, the items preferred under the MI method have large discrimination and large upper asymptote. Note that $(3 - \sqrt{9 - 8d_{MI}^{(t+1)}})/4$ increases in $d_{MI}^{(t+1)}$. So does $f(d_{MI}^{(t+1)})$ as shown earlier. Therefore, large upper asymptote tends to increase the step size in Equation 18; that is, when the MI method is used, its preference of high- d items would lead to a big reduction in the ability estimate in the presence of slipping, on top of the effect of large- a .

However, the product of $f(d_{MI}^{(t+1)})(3 - \sqrt{9 - 8d_{MI}^{(t+1)}})/4$ ranges from $(0, 1/2)$. Therefore, just by including an upper asymptote between 0 and 1, the step size can be reduced to $f(d_{MI}^{(t+1)})(3 - \sqrt{9 - 8d_{MI}^{(t+1)}})/2$ times its original size. Here, original size refers to the step size when $d = 1$, that is, $1/2$. For $d_{MI}^{(t+1)} = .75$, the step size reduces to 53.6% of its original size, when the examinee answers the $(t + 1)$ th item right by lucky guessing. In the extreme case, when $d_{MI}^{(t+1)} = 0$, the step size becomes 0. The solid line in the right panel of Figure 1 shows how the relative step size changes with d , when items are chosen with the MI method.

(Relative) step size under b-matching under the 3PLMU. If the item pool is rich enough, under the b -matching method $b^{(t+1)} = \hat{\theta}^{(t)}$, and the probability of answering the next item correctly is $d_{BM}^{(t+1)}/2$. Correspondingly, the step size in the presence of slipping is,

$$\frac{a_{BM}^{(t+1)}}{I^{(t+1)}(\theta^*)} f(d_{BM}^{(t+1)}) \left(\frac{d_{BM}^{(t+1)}}{2} \right). \quad (19)$$

Equation 19 also shows that having an upper asymptote smaller than 1 rather than fixing it at 1 makes the step size smaller, because $f(d_{BM}^{(t+1)})(d_{BM}^{(t+1)}/2)$ is an increasing function in the upper asymptote. The step size can be reduced to $f(d_{BM}^{(t+1)})(d_{BM}^{(t+1)})$ times its original size. Here, original size refers to the step size when the upper asymptote is fixed at 1 (as in the usual IRT models), which is $1/2$. When $d_{BM}^{(t+1)} = .75$, the step size is reduced to 63.4% of its original size.

Asymmetry: Early guessing versus early slipping under the 3PLMU. Analogous to the results regarding the lower asymptote, having an upper asymptote parameter between 0 and 1 helps reduce the step size to a larger degree under the MI method than the b -matching method, as shown by the case when $c = .25$. This is because other things being equal, the ratio of step size under the MI versus the b -matching strategy is $(3 - \sqrt{9 - 8d_{MI}^{(t+1)}})/2d_{BM}^{(t+1)}$, which is smaller than 1 when $0 < d < 1$. The dotted curve in the right panel of Figure 1 shows how the relative step size responds to d when the b -matching method is used for item selection. As expected, the dotted curve dominates the solid curve in the right panel as well, for $\forall d \in (0, 1)$. Interestingly, the left and right panels in Figure 1 are mirror images of each other, that is, the relative step size is numerically the same for $c = 1 - d$. For example, when $c = .25$ or $d = .75$, the relative

step sizes are both .536. Simply put, the upper asymptote and lower asymptote have complementary effects.

Because of such complementary effect, the positive bias related to early guessing is expected to be larger in magnitude than the negative bias related to early slipping under the 3PLMU. The step size under the MI method given an incorrect answer from slipping is provided in Equation 18. Meanwhile, the step size when a correct answer is given from guessing is,

$$\frac{a_{MI}^{(t+1)}}{I^{(t+1)}(\theta^*)} f(d_{MI}^{(t+1)}) \left(\frac{1 + \sqrt{9 - 8d_{MI}^{(t+1)}}}{4} \right). \quad (20)$$

The ratio of the step size under slipping versus guessing is $(3 - \sqrt{9 - 8d_{MI}^{(t+1)}})/(1 + \sqrt{9 - 8d_{MI}^{(t+1)}})$, which is smaller than 1 for $\forall d_{MI}^{(t+1)} < 1$ (proof omitted). This formally shows that the loss of early slipping is not as big as the gain of early guessing given an upper asymptote smaller than 1, as the upper asymptote explicitly accounts for slipping.

If the *b*-matching method is used for item selection, the step size under slipping is given in Equation 19. The step size under guessing is $(a_{BM}^{(t+1)}/I^{(t+1)}(\theta^*))f(d_{BM}^{(t+1)})(1 - (d_{BM}^{(t+1)}/2))$. The ratio is, therefore, $(d_{BM}^{(t+1)})/(2 - d_{BM}^{(t+1)})$, which monotonically increases in $d_{BM}^{(t+1)}$. So the ratio is smaller than 1 for $\forall d_{BM}^{(t+1)} \in (0, 1)$. This formally shows that the benefit of early lucky guesses trumps the loss of early slips given an upper asymptote smaller than 1 under the *b*-matching method. The ratio of loss to gain is $(3 - \sqrt{9 - 8d_{MI}^{(t+1)}})/(1 + \sqrt{9 - 8d_{MI}^{(t+1)}})$ and $(1 - (d_{BM}^{(t+1)}/2))$ for the MI and the *b*-matching method, respectively.

In summary, the upper asymptote and lower asymptote have complementary effects. Under the 4PLM, the positive bias is smaller than that under the 3PLMU, and the negative bias is smaller than that under the 3PLM. It is also expected that given any IRT model, the *a*-stratified design helps reduce the bias caused by early slipping or guessing.

Conclusion and Discussion

By re-examining and extending the theoretical results in Chang and Ying (2008), the authors show in this article analytically that under both the MI method and the *b*-matching method, large discrimination parameter has a magnifying effect on the step size. The adverse effect of early slipping or guessing is therefore exacerbated under the MI method, which favors highly discriminating items early in the test. The adverse effect should be smaller under the *b*-matching method, which does not show favoritism to highly discriminating items. More interestingly, the preference of the MI method to the low-*c* and high-*d* items also exacerbates the adverse effect of early slipping or guessing.

Analytical results provided in this paper further prove that having the upper or lower asymptote parameters rather than fixing them at 1 or 0 (i.e., essentially leaving them out of the model) helps make the step size smaller when early slipping or guessing happens, respectively. The *relative* step sizes are derived under both the MI and the *b*-matching method. Notably, the reduction in step size is more pronounced under the MI method than under the *b*-matching method.

Such analytical results indicate that the effect of early misbehaviors in CAT can be moderated by the inclusion of proper parameters in the IRT models. In fact, models explicitly accounting for both slipping and guessing are not uncommon, for example, the deterministic inputs, noisy “and” gate (DINA) and deterministic inputs, noisy “or” gate (DINO) model in the literature of cognitive diagnosis (Rupp, Templin, & Henson, 2010). The use of the 4PLM in operational CAT, however, requires further investigation on the fit of the 4PLM to real testing data,

especially pretesting data used for item calibration, as well as the precision of the asymptote parameter estimates. As noted by the reviewers, obtaining precise estimates of the c parameter in the 3PLM is known to be a challenge unless the sample size is large and the sample covers the lower end of the ability continuum adequately. The same applies to the estimation of the upper asymptote parameter d . The sample size required to accurately calibrate the 4PLM is even larger and there needs to be enough slipping happening in the data. Recent discussions of the calibration of the 4PLM such as Loken and Rulison (2010) adopted Markov Chain Monte Carlo (MCMC) method, which means priors are imposed on these parameters. It is still an open question how to pick appropriate choices of the priors and the magnitude of influence of informative priors on the estimation outcomes. Overall, this article provides required technical details to implement a CAT based on the 4PLM when item parameters are available, and show theoretical advantages of including proper asymptote parameters in the model to combat guessing or slipping. However, the authors advise readers to exercise caution when they decide on which IRT model to use. Model fit and precision of parameter estimates need to be evaluated.

In addition, the analytical results are derived under the MI method and b -matching method. When other mechanisms are introduced in the item selection of CAT, for example, certain exposure control strategies, the step sizes in Equations 12, 13, 18, and 19 will no longer apply and neither will the relative step sizes. But the general formula (Equations 6, 7, and 16) are still valid. If the add-on mechanism alleviates the reliance on highly discriminating items early in the test, the adverse effect of early guessing or slipping can be reduced compared with MI method. For example, the a -stratified design should help reduce the negative bias of early slipping under the 3PLM or the positive bias of early guessing under the 3PLMU, while leading to much more balanced item pool usage than the MI method. The a -stratified method was initially proposed to control the exposure of highly discriminating items and promote the use of low-discrimination items in CAT. It groups items into several strata according to their discrimination parameters, from low- a stratum to high- a stratum. In the early stage of a CAT, items can only be chosen from the low- a stratum and gradually moves on to higher strata (Chang, Qian, & Ying, 2001; Chang & Ying, 1999). Such “ascending a ” principle is shown effective for balancing item exposure in Chang and Ying (2008) and Cheng, Chang, Douglas, and Guo (2009). Because the step size is proportional to the a -parameter under the 4PLM and the smaller models it subsumes, using low- a items early in the test helps reduce the effect of early slipping and early guessing.

One last thing to emphasize is that the analytical results in this article apply to maximum likelihood ability estimates. Rulison and Loken (2009) used simulation study to show how the upper asymptote helps reduce negative bias associated with early slipping, by contrasting the posterior distribution of ability given the same response pattern under the 3PLM versus under the constrained 4PLM model by Barton and Lord (1981). The analytical results in this article can be extended to Bayesian estimates of ability, for example, maximum a posteriori (MAP) estimates. This will be pursued in the future.

Acknowledgment

The authors would like to thank the action editor, Dr. Dan Bolt, and two anonymous reviewers for their insightful suggestions and comments.

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: The authors thank the Institute for Scholarship in the Liberal Arts (ISLA) at the University of Notre Dame for a Small Grant for Research and Creative Work awarded to the first author

Supplemental Materials

The online appendices are available at <http://apm.sagepub.com/supplemental>.

References

- Barton, M. A., & Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic item-response model* (Tech. Rep. RR-81-20). Princeton, NJ: Educational Testing Services.
- Birnbaum, A. (1968). Some latent ability models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 392-479). Reading, MA: Addison-Wesley.
- Chang, H.-H., Qian, J., & Ying, Z. (2001). Alpha-stratified multistage computerized adaptive testing with beta blocking. *Applied Psychological Measurement*, 25, 333-341.
- Chang, H.-H., & Ying, Z. (1999). Alpha-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23, 211-222.
- Chang, H.-H., & Ying, Z. (2008). To weight or not to weight? Balancing influence of initial items in adaptive testing. *Psychometrika*, 73, 441-450.
- Cheng, Y., Chang, H.-H., Douglas, J., & Guo, F. (2009). Constraint-weighted a-stratification for computerized adaptive testing with nonstatistical constraints. *Educational and Psychological Measurement*, 69, 35-49.
- Cheng, Y., & Keng, L. (2009). Computerized adaptive testing in criterion-referenced testing. In E. Smith & G. Stone (Eds.), *Applications of Rasch measurement in criterion-reference testing: Practice analysis to score reporting*. (pp. 458-487). Maple Grove, MN: JAM Press.
- Embretson, S. E., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.
- Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement*. Homewood, IL: Dow Jones-Irwin.
- Kamat, S. (2013). *GMAT exam—How important are the first few questions*. Retrieved from www.mbacrystalball.com/blog/2013/07/23/gmat-exam-questions-timing/
- Koprince, S. (2012). *Debunking the "first 7 questions" myth*. Retrieved from <http://www.manhattangmat.com/blog/index.php/2012/03/14/debunking-the-first-7-questions-myth/>
- Liao, W.-W., Ho, R.-G., Yen, Y.-C., & Cheng, H.-C. (2012). The four-parameter logistic item response theory model as a robust method of estimating ability despite aberrant responses. *Social Behavior and Personality*, 40, 1679-1694.
- Loken, E., & Rulison, K. (2010). Estimation of a four-parameter item response theory model. *British Journal of Mathematical and Statistical Psychology*, 63, 509-525.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillside, NJ: Lawrence Erlbaum.
- Magis, D. (2013). A note on the item information function of the four-parameter logistic model. *Applied Psychological Measurement*, 37, 304-315.
- McDonald, R. P. (1967). *Nonlinear Factor Analysis* (Psychometric Monograph No. 15). Richmond, VA: Psychometric Corporation. Retrieved from <http://www.psychometrika.org/journal/online/MN15.pdf>.
- Reise, S. P., & Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items? *Psychological Methods*, 8, 164-184.
- Rulison, K., & Loken, E. (2009). I've fallen and I can't get up: Can high-ability students recover from early mistakes in CAT? *Applied Psychological Measurement*, 33, 83-101.

- Rupp, A., Templin, J., & Henson, R. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York: Guilford Press.
- Sackmann, J. (2010). *How important are the first 5-10 questions of each GMAT section?* Retrieved from <http://www.gmathacks.com/cat-strategy/how-important-are-the-first-5-10-questions-of-each-gmat-section.html>
- Waller, N. G., & Reise, S. P. (2009). Measuring psychopathology with non-standard IRT models: Fitting the four parameter model to the MMPI. In S. Embretson & J. S. Roberts (Eds.), *New directions in psychological measurement with model-based approaches* (pp. 147-173). Washington, DC: American Psychological Association.
- Yen, Y.-C., Ho, R.-G., Laio, W.-W., Chen, L.-J., & Kuo, C.-C. (2012). An empirical evaluation of the slip correction in the four parameter logistic models with computerized adaptive testing. *Applied Psychological Measurement*, 36, 75-87.