9

A HYBRID Model for Test Speededness

Keith A. Boughton[1] and Kentaro Yamamoto[2]

[1] CTB/McGraw Hill
[2] Educational Testing Service

9.1 Introduction

Assessing speededness by simple approaches such as counting the number of missing responses near the end of a test is often inadequate because many examinees switch to a guessing or random response strategy as the testing time limit approaches. Parameter estimation within item response theory (IRT) can be greatly impacted by speededness; thus, it is crucial to assess how much speededness a test may possess (Oshima, 1994). It is also critical to correct the item-parameter estimates that may have been affected by this end-of-test speededness. Examinees who switch to random responses at the end of the test, in terms of the underlying response processes, expose a very different behavior when responding to an item when compared to examinees who try to solve the item using their cognitive skill set. In order to account for these different types of response behaviors, a HYBRID model was proposed by Yamamoto (1989) and later extended to assess test speededness more directly (Yamamoto, 1990, 1995).

It is important to note that there are many reasons why one of the key assumptions of IRT, namely that of conditional independence, may fail. Speededness is one such case, and in particular, this application of the HYBRID model addresses a specific type of speededness that will be later elaborated on. This research will show how the HYBRID model can detect examinees who have switched to a random response strategy, thereby eliminating the noise caused by end-of-test speededness, which should result in more accurate IRT parameter estimates for those end-of-test items.

9.2 Purpose and Method

This chapter will first explicate the HYBRID model, its development, and parameter estimation. The second section will demonstrate, using real data with quasi-experimental controls, the HYBRID model's accuracy and efficacy

for assessing the amount of speededness and reducing its effects on item-parameter estimation. A writing assessment with 45 multiple-choice items was shortened to 38 items due to known end-of-test speededness and then readministered. The last five items of the 45-item test were placed in the middle of the shortened 38-item version, thus creating a quasi-experimental condition in which the item parameters from the middle of the 38-item test, which should not have been affected by speededness, will be used as the estimated "true" parameters. The HYBRID model item-parameter estimates for the last five items of the longer 45-item version will be compared to these "true" parameters.

9.3 The HYBRID Model

The original HYBRID model, proposed by Yamamoto (1987, 1989), was specifically developed in order to incorporate cognitive structure into the IRT methodology of that time, which up to that point, was mostly used for the scaling and reporting of scores from large-scale assessments. Yamamoto (1989) acknowledges that a multidimensional IRT (MIRT) model could be employed; however, he cautions against the use of the compensatory MIRT model, since assessments may not involve compensatory abilities. He also points out that, "the notion of single-event learning cannot be incorporated easily into a purely continuous model" (p. 4).

Yamamoto (1990) later extended the HYBRID model for diagnosing test speededness. This psychometric approach to speededness has made significant advances in this area by combining a latent-class (LC) model with an IRT model strategy. This HYBRID model has been studied through several simulations (Boughton & Yamamoto, 2004; Yamamoto, 1990, 1995; Yamamoto & Everson, 1995, 1997). As with any model, however, more research is needed in order to securely support and demonstrate its appropriateness and utility, especially with the use of real data, since simulations cannot model actual human response behavior.

Classical test theory (CTT) and item response theory (IRT) each describe the behavior of examinees based on a single model, whereas the HYBRID psychometric approach (Yamamoto, 1989) utilizes two models in the detection of speededness. That is, subsets of examinee response patterns are modeled by a discrete latent-class model (i.e., multinomial independent class), with the remaining responses modeled by an IRT model (Yamamoto & Everson, 1995, 1997). In contrast to finite-mixture-distribution IRT models that assume the same parametric model—with different parameter vectors—in each of the mixing components, HYBRID models assume different model structures in each mixture component. It is important to note that the HYBRID model does not necessarily have only two classes, but is implemented by assuming many classes with restrictions imposed across classes, each defined by a switch point in the item sequence. The HYBRID model can estimate the

point in an assessment at which each examinee has switched from an ability-based response strategy to a guessing or random-response strategy. Thus, the HYBRID model provides an index to help set test lengths appropriate to the time-constraint allocations, as well as to ascertain the differential speededness for any subgroup population (Boughton et al., 2004; Yamamoto & Everson, 1995).

9.4 HYBRID Model and Parameter Estimation

The HYBRID model estimates both person and item parameters along with the parameters that define the distribution of examinees switching from an ability-based to a random-response strategy. The HYBRID model assumes that any examinee who switches to a random response strategy has conditional probabilities that are independent of their proficiency level for the remaining items. Every examinee's response can be modeled either by a continuous unidimensional IRT model or an LC model, and conditional independence holds, given an examinee's proficiency and strategy. The following function expresses the likelihood of a correct response on an item i given the three assumptions above:

$$p\left(x_i = 1 | \, \theta, \beta_i, k\right) = \left(1 + \exp\left(\theta - b_i\right)\right)^{m_{ik}} c_i^{m_{ik}+1}, \tag{9.1}$$

where k indicates the last item answered under the IRT model; $M_{ik} = -1$, when $i \leq k$ and $M_{ik} = 0$, when $i > k$. x_i is a dichotomous response (i.e., 0/1) on item i; β_i represents the item difficulty parameter; θ is the examinee ability parameter; and c_i is the expected proportion correct under a patterned or random response strategy. Equation 9.1 gives the conditional probability of a response x_i, given θ, item parameters β_i, and strategy switch point k. Specifically, this function specifies that an IRT model holds until a random response occurs, with a constant conditional probability holding for the remaining random responses (Yamamoto, 1995).

The likelihood of observing a response vector x_v, given θ_v, when switching from an ability-based solution to a random-response strategy on item k_v is

$$P\left(x_v | \, \theta_v, B, k_v\right) = \prod_{i=1}^{k_v} P\left(\theta_v, \beta_i\right)^{x_{iv}} Q\left(\theta_v, \beta_i\right)^{1-x_{iv}} \prod_{i=k_v+1}^{I} c_i^{x_{iv}} \left(1 - c_i\right)^{1-x_{iv}}. \tag{9.2}$$

The marginal probability of observing x_v given model parameters B is

$$P(\, x_v | \, B) = \sum_{k}^{L} \int_{\theta} P(x_v \, | \theta, B, k) \, f(\theta \, | k) d\theta f(k) \; , \tag{9.3}$$

where $f(\theta \, | k)$ is the conditional probability of θ given a switch point k, and $f(k)$ is the marginal distribution of the strategy-switching population.

The joint likelihood of parameters given the observed response matrix $X = (x_1, x_2, \ldots, x_v)$ from a total of V examinees is

$$L\left(B\,|X\right) = \prod_{v=1}^{V} P\left(x_v\,|B\right).$$

The IRT item parameters can be estimated to maximize the above marginalized likelihood function using an iterative method, such as the Newton–Raphson (N-R) method. The N-R method can be described as $P^{n+1} = P^n - D_2^{-1} * D_1$, where P^{n+1} is a vector of parameters updated from P^n by a certain amount designated by the function D_2 (matrix of second derivatives) and D_1 (vector of first derivatives). However, D_2 can be quite large and the off-diagonal elements need not be zero. Consequently, a full implementation of the N-R method would be too great a computational burden. Bock & Aitkin (1981) advanced the idea of using the EM algorithm (Dempster et al., 1977) in the area of IRT parameter estimation. Within the EM algorithm, the continuous distribution of theta (i.e., the ability parameter) is approximated by a discrete distribution, in order to facilitate the numerical integration over the range of the latent-variable theta. With respect to u, a model parameter including either an item parameter or a probability of the discrete ability density, the first derivative of the log-likelihood of the above function can be expressed as

$$\frac{\partial \ln L\left(B\,|X\right)}{\partial u} = \sum_{v=1}^{V}\sum_{k=1}^{I}\int_{\theta}\frac{\partial P\left(x_v\,|\theta, B, k\right)}{\partial u}\frac{f\left(\theta\,|k\right)f\left(k\right)}{P\left(x_i\,|B\right)}d\theta.$$

Followed by the application of the empirical Bayes method and approximation of integration by summation denoted by q-quadrature points and $A\left(\theta_q\,|k\right)$ as defined as conditional weights approximating $f\left(\theta_q\,|k\right)$, the above equation for a parameter u_i can be written as

$$\frac{\partial \ln L}{\partial u_i} = \sum_{k}\sum_{q}\frac{A\left(\theta_q\,|k\right)}{P_{ik}\left(\theta_q\right)Q_{ik}\left(\theta_q\right)}\frac{\partial P_{ik}\left(\theta_q\right)}{\partial u_i}\sum_{v=1}^{V}\left[x_{iv} - P_{iv}\left(\theta_q\right)\right]f\left(k\right)P_i\left(\theta_q\,|x_v, k\right).$$

The right side of the above equation can be rewritten as follows, since x_{iv} is either 1 or 0:

$$\sum_{k}\sum_{q}\frac{1}{P_{ik}\left(\theta_q\right)Q_{ik}\left(\theta_q\right)}\frac{\partial P_{ik}\left(\theta_q\right)}{\partial u_i}f\left(k\right)\left(R_{iqk} - P_{ik}\left(\theta_q\right)N_{iqk}\right),$$

where

$$R_{iqk} = \sum_{v} x_{iv}\frac{P\left(x_v\,|\theta_q, B, k\right)A\left(\theta_q\,|k\right)}{P\left(x_v, B\right)},$$

$$N_{iqk} = \sum_{v}\frac{P\left(x_v\,|\theta_q, B, k\right)A\left(\theta_q\,|k\right)}{P\left(x_v, B\right)},$$

and

$$\frac{\partial P_{ik}(\theta_q)}{\partial a_i} = D(\theta_q - b_i) P_{ik}(\theta_q) Q_{ik}(\theta_q),$$
$$\frac{\partial P_{ik}(\theta_q)}{\partial b_i} = -DaP_{ik}(\theta_q) Q_{ik}(\theta_q).$$

The matrix of second-order derivatives can be expressed as follows:

$$\frac{\partial^2 \ln L}{\partial a_i^2} = D^2 \sum_k \sum_q f(k) (\theta_q - b_i)^2 N_{iqk} P_{ik}(\theta_q) Q_{ik}(\theta_q),$$
$$\frac{\partial^2 \ln L}{\partial b_i^2} = -b^2 \sum_k \sum_q a_i^2 N_{iqk} P_{ik}(\theta_q) Q_{ik}(\theta_q),$$
$$\frac{\partial^2 \ln L}{\partial a_i \partial b_i} = D^2 \sum_k \sum_q a_i (\theta_q - b_i)^2 N_{iqk} P_{ik}(\theta_q) Q_{ik}(\theta_q).$$

Once item parameters are estimated, estimation of an examinee's proficiency can be carried out using one of several existing methods, such as the maximum likelihood method (MLE), Bayes modal estimates (MAP), or expected a posteriori (EAP). The MLE ability estimation is described by Lord (1980), and MAP and EAP are both described by Bock & Aitkin (1981).

Prior distributions for the item parameters, proficiency, and switching population distributions can be used during the maximization phase. For example, item parameters can be assumed to be drawn from a particular distribution, and, therefore, updating parameters would be constrained to meet that particular distribution. Likewise, the proficiency distribution may be assumed as a normal distribution at each switching point, including the last item. In addition, $E(\theta | k)$ may be constrained to have a specific functional form in relation to the value of k (Yamamoto, 1995). The HYBRID model parameters for the speededness model can be estimated using the HYBILm software program (Yamamoto, 1990).

9.5 Results

The 44-item writing assessment was shortened to 38 items after the last five items were found to be greatly impacted by speededness (i.e., student reported speededness). These five items were then repositioned into the middle of the 38-item form, giving us the opportunity to demonstrate how well the HYBRID Rasch model (RM) can recover the "true" parameters (i.e., the parameter estimates obtained when calibrated in the unaffected portion of the shortened 38-item test). The parameters of the five items in the middle of the shortened 38-item test will be considered the "true" parameters, and the comparisons

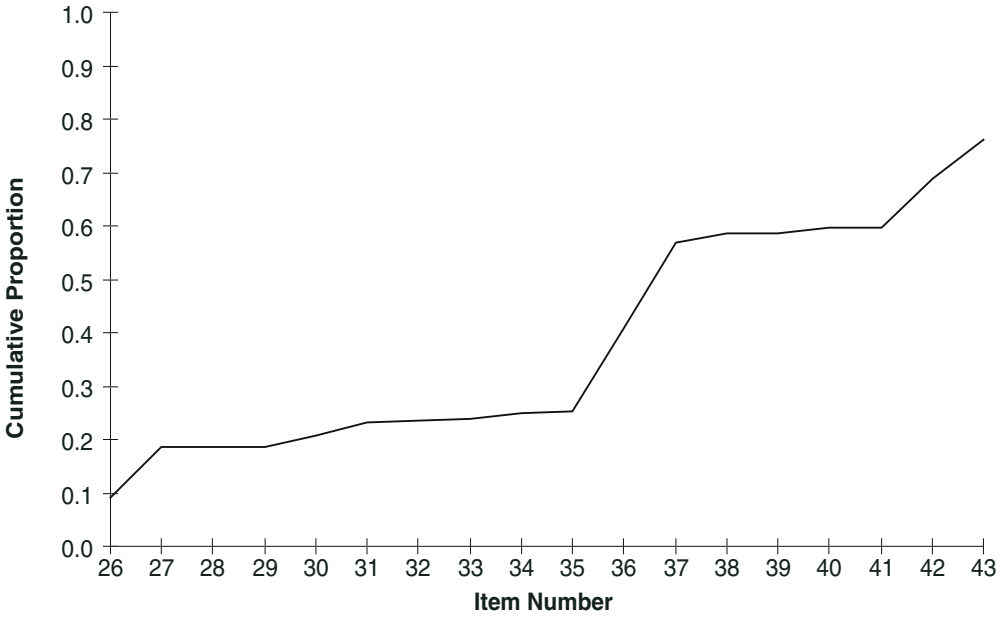

Fig. 9.1. Cumulative switching proportions across items 26–43 in the 44-item test

will be made between these five-item-parameter estimates and the estimates obtained from the end of the 44-item calibration.

Figure 9.1 displays the cumulative proportion of examinees switching from an ability-based strategy to a random-response strategy across the last 18 items. The x-axis represents the item number and the y-axis is the cumulative proportion of examinees switching strategies. As seen from the figure, this test is speeded, with over 50% switching to a random-response strategy starting at item 37. Figure 9.2 displays the cumulative proportion switching across the last 17 items of the shortened 38-item form. The proportion of switchers is considerably lower. However, there is still over 20% switching over the last four items. The HYBRID model can be used as a tool to help identify how short a test needs to be in order to give all examinees the opportunity to show their true abilities fairly. Given the switching information from Figure 9.1, it would seem reasonable to shorten the test to a length of 35 items; however, the test was only shortened to a length of 38 items, given reliability predictions and time-per-item estimations. Note that the switching proportions would suggest that the test should be shortened to 35 items, since we observe approximately 20% switching on that item. Although the 20% criterion is a somewhat arbitrary bound, given the authors' experiences with speeded assessments and their effects on item-parameter estimation, it seems a good rule of thumb. Of course, it would be more desirable to have 0%, at least for assessments in which speed of answering is not the intended construct; however, this may not be realistic. Thus, it is the impact on the item-parameter estimates that will be the defining factor in this research. It can be seen in Figure 9.2 that

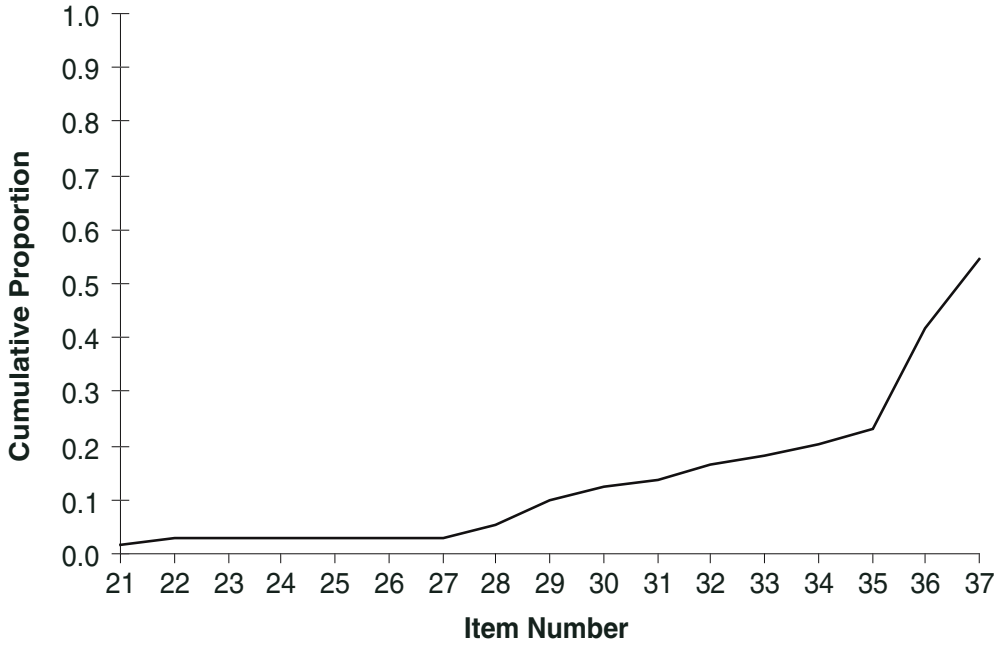

Fig. 9.2. Cumulative switching proportions across items 21–37 in the 38-item test

the shortened test was not short enough, and had it been shortened to 35 items, as the HYBRID model suggests, then the 20% criterion would most likely have been met when the test was readministered. However, it could be that, no matter how short an assessment is, there will always be examinees who cannot estimate how much time it will take to complete the test.

Figure 9.3 shows the item-characteristic curves for the five items that were removed from the end of the 44-item test and moved to the middle of the 38-item test. The actual position of item 39 in the 38-item version is (22), 40 (23), 42 (24), 43 (25), and 44 (26). Each of the five graphs has three ICCs; the "true" ICCs (i.e., recalibrated in the middle of the 38-item test), the Rasch ICCs, and the HYBRID ICCs, both calibrated in the 44-item-test version and then scaled using a Stocking & Lord (1983) transformation to the 38-item-test scale, using the first 21 nonspeeded items in both tests. All items, except for item 22, were biased when the Rasch model was used alone (i.e., items appeared more difficult). However, the HYBRID RM produced corrected item parameters that were consistent with the "true" parameters, with a slight overcorrection for items 24, 25, and 26 (i.e., the item appeared slightly easier).

Figure 9.4 displays the speeded characteristic curves. The x-axis is the ability metric, with the y-axis being the expected true score for the five items presumed speeded. The impact of the bias in the expected score would be about 0.5 for the middle of the ability range. The "Rasch-only" model is

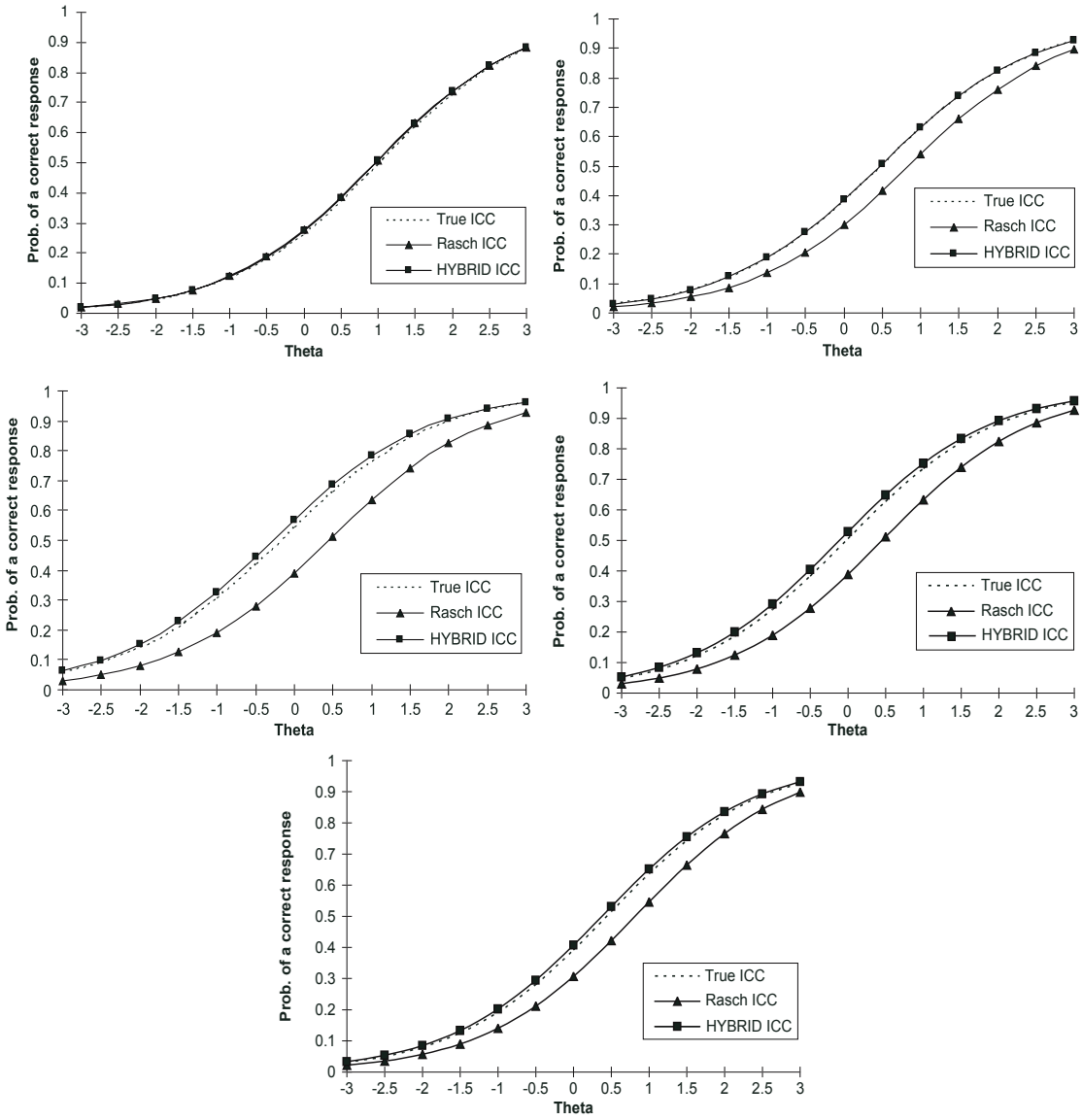

Fig. 9.3. Item characteristic curves for "true," Rasch-only, and HYBRID-Rasch for Items 22, 23, 24, 25, and 26, from left to right and down

biased and would result in a lower-ability expected score. The HYBRID model recovered the "true" five item parameters.

Figure 9.5 displays the entire 38-item-test test characteristic curve TCC, for the "true," the Rasch-only, and the HYBRID TCC. The TCC is recovered when the HYBRID model is used, while the Rasch-only is biased.

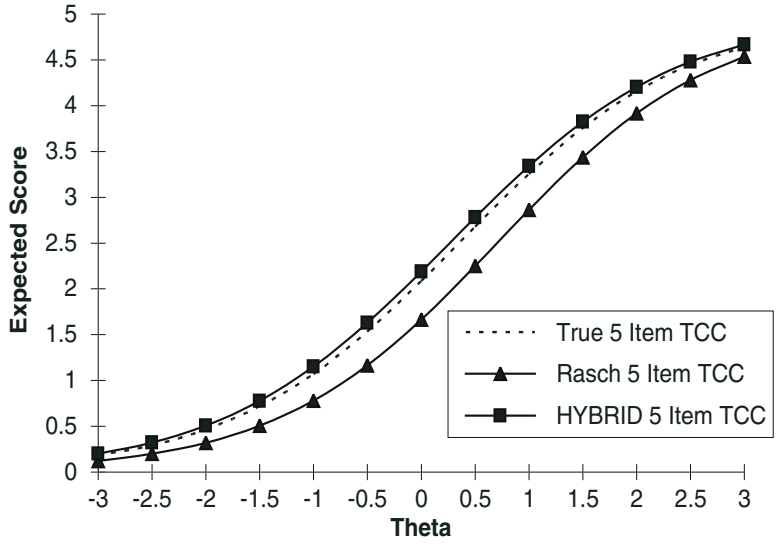

Fig. 9.4. Speeded-section characteristic curves for "true," Rasch-only, and HYBRID-Rasch

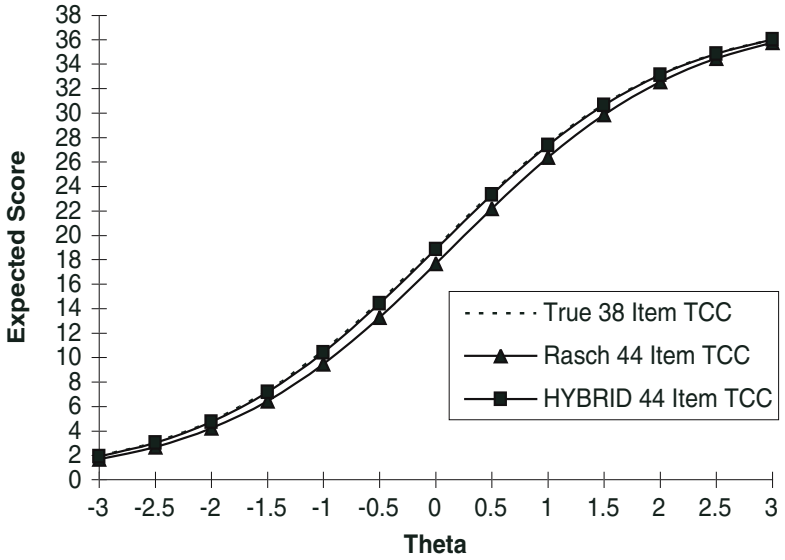

Fig. 9.5. Test characteristic curves for "true," Rasch-only, and HYBRID-Rasch

9.6 Discussion and Conclusion

The purpose of this study was to examine an estimation method that incorporates two distinct models for each examinee in the detection and then modeling of speededness. This research has demonstrated the HYBRID model's utility and appropriateness using real data with quasi-experimental controls.

The 44-item test was found to be speeded, with over 20% switching to a random response pattern on item 36. However, even the shortened 38-item version shows over 20% switching behavior before reaching the last item. These results suggest that even when examinees are given more time per item, some examinees do not pace themselves appropriately and thus fail to reach the end of the test using an ability-based response strategy. When not accounting for speededness, parameter-estimation bias was found in four of the five items studied, with the Rasch-only model overestimating the difficulty of the items. The HYBRID RM corrected all of the parameter estimates, although it slightly underestimates the item difficulty for some of the items. However, at the TCC level, the HYBRID RM matches the "true" TCC, while the Rasch-only model results in a biased TCC. These results suggest that the HYBRID model improves item-parameter estimation for speeded items located near the end of a test. These improvements coincide with the proportion of examinees switching to a random-response strategy on each form.

The HYBRID model provides a method that can reduce the effects of speededness on IRT item and ability parameters, while also mapping item/examinee switch behavior for tests with speededness. However, the HYBRID model does not work well for all testing situations. For example, if examinees responded randomly at the beginning of a test, then the current model would not be appropriate. The HYBRID model also does not work well for tests that have items ordered from easiest to most difficult, because low-ability examinees will have response patterns similar to examinees switching to a random-response strategy (Yamamoto & Everson, 1995). The tests presented in this study did not have any of these limitations. Ironically, as is the case for many studies, this research's strength is also its weakness. The application to real-world data with quasi-experimental controls is paramount in illustrating the HYBRID model's utility and appropriateness. However, the accuracy of the parameter estimates are judged in comparison with parameters that are estimates in and of themselves. In addition, position effects may hamper direct comparison between the long and shortened test length item parameters, although this was not apparent with these results. It is extremely important to ensure that test length or time is appropriate when a test's construct of interest does not include the speed with which each student answers. Searching for not-reached items at the end of a test, especially for examinees who randomly fill in unanswered responses, may not prove beneficial. In these cases, the HYBRID RM proposed here can aid test developers in setting appropriate test lengths (i.e., using the cumulative proportion switching), and/or correct any speededness-induced bias for end-of-test items.