An item response theory model of matching test performance

Matthew D. Zeigenfuse a, William H. Batchelder b,1, Mark Steyvers b,*

a Quartet Health, United States of America
b Department of Cognitive Sciences, University of California, Irvine, United States of America

ARTICLE INFO

Article history:
Received 2 February 2019
Received in revised form 2 January 2020
Accepted 24 January 2020
Available online xxxx

Keywords:
Matching tests
Rasch model
Item response theory
Bayesian inference
Markov Chain Monte Carlo

ABSTRACT

In a matching test, a test taker is presented with a list of test items and a list of response alternatives and asked to match each response alternative with a test item. The response alternatives can be given as a response to at most one test item. As a result, the response a test taker offers to one test item depends on his or her responses to all of the other test items. This violates the “local independence” assumption underlying most existing item response theory (IRT) methods, such as the Rasch model. Here we develop a framework for extending dichotomous IRT models to account for test taking behavior on matching tests. This model separates an individual's knowledge of the correct responses to the items of a matching test from his or her responses to those items. In addition to developing the matching framework, we derive a number of important properties, including its item response function and score distribution. Finally, we demonstrate through an empirical example that our matching test framework provides a good account of behavior on matching tests.

© 2020 Elsevier Inc. All rights reserved.

1. Introduction

Cultural Consensus theory (CCT) is one of the key contributions of Bill Batchelder and honoree of the special volume. CCT can be used to derive the consensus knowledge of a group or population as well as participants’ degree of adherence to consensus knowledge. Initially, CCT was applied to true–false questions (Batchelder & Romney, 1988; Romney, Batchelder, & Weller, 1987) but more recently, it has been extended to ordinal data (Anders & Batchelder, 2015) as well as continuous responses (Anders, Oravecz, & Batchelder, 2014). In addition, CCT has also formed the basis of cognitive models of wisdom of crowds (Lee, Steyvers, De Young, & Miller, 2012; Lee, Steyvers, & Miller, 2014). Bill Batchelder had a longstanding interest in extending cultural consensus theory (CCT) models to matching tests, which are simple and practical, and are often included in attainment tests, but they also violate an important assumption of CCT, local independence.

Fig. 1 offers an example of a test comprised of matching items. Generally, a matching test will contain one or more sets of items like Fig. 1; however, for the moment we will consider a short test comprised of only a single set of matching items. In this example matching test, each test taker assigns a single philosopher’s last name in the left-hand column to exactly one response alternative from the first names in the right-hand column. The response alternatives that are not the correct response to any item will be called “distractors”, since their purpose is to distract test takers from the correct answer when guessing. Though not as widely used as true–false or multiple choice questions, matching tests like Fig. 1 can nonetheless be found in various contexts, such as attainment tests (e.g., Aiken, 1997; Gronlund, 1998; Heywood, 1977; Miller, Linn, & Gronlund, 2009; Nitko & Brookhart, 2010), educational research (e.g., Bean, Searles, Singer, & Cowen, 1990; Craig, Gholson, & Driscoll, 2002; Moreno & Mayer, 1999; Popp, 1964) and psychometric measurement (e.g., McDonald, 2013; Osterweil, Mulford, Syndulko, & Martin, 1994). They are employed by classroom teachers for measuring learning and retention (Haynie, 2003; Moore, 2001). Studies have also found that test takers score higher on matching tests (Benson, 1981; Benson & Crocker, 1979), perhaps owing to lower test taker anxiety (Shaha, 1984).

In this paper, we develop an item response theory (IRT) model for matching tests. IRT and CCT are closely related, and much of the model development was based on earlier (unpublished) work between the co-authors on a CCT model for matching tests.

Item response theory (IRT) is a large class of statistical models for measuring a latent, often unidimensional, trait or ability using a test or questionnaire. IRT models have been developed for a...
Match each philosopher’s last name to their first name.

| Rousseau | Jean-Jacques |
| Sartre   | John         |
| Hobbs    | Jean-Paul    |
| Nietzsche| Thomas       |
| Locke    | Franz        |
|          | Friedrich    |
|          | Philip       |

Fig. 1. Sample set of matching items. The items are the last names in the left column and the possible responses are the first names in the right columns. The correct responses to the items are “Jean-Jacques”, “Jean-Paul”, “Thomas”, “Friedrich” and “John”.

variety of response formats, including true–false responses (Birnbaum, 1968; Rasch, 1960), nominal responses (Bock, 1972), multiple choice responses (Thissen & Steinberg, 1984) and ordered responses (Andrich, 1978; Masters, 1982; Samejima, 1969). In general, these models specify the likelihood of the latent trait or ability given the individual’s responses to a test with M items, allowing it to be inferred from his or her responses to the test items using statistical methods such as marginal maximum likelihood estimation (e.g., Molenaar, 1995) or Bayesian inference (e.g., Fox, 2010).

Despite the breadth and applicability of matching tests, no IRT models exist for inferring each test taker’s ability from the pattern of test responses of a matching test. The difficulty of developing such a model is two-fold. Standard IRT models, like the Rasch and two-parameter logistic (2PL) models assume responses are locally independent for all pairs of subjects and items. However, responses to matching items are not locally independent. Once a response is given for one item it cannot be given as a response to another. For example, if a test taker correctly provides the response “Jean-Jacques” to the item “Rousseau”, he or she cannot also provide “Jean-Jacques” as a response to the item “Sartre”, making its conditional probability zero. Alternatively, if the test taker incorrectly provides the response “John” to “Rousseau”, the response “Jean-Jacques” remains as a possible response for “Sartre”, making its conditional probability non-zero. Thus, the response probabilities to “Sartre” depend conditionally on the test taker’s response to “Rousseau”, meaning the two responses are not independent.

The second difficulty for developing an IRT model for matching tests is that test takers’ response patterns only offer imperfect information about whether they know the correct response to an item. For example, a test taker who knows that the correct answers to “Rousseau”, “Sartre”, “Hobbs” and “Locke” are, respectively, “Jean-Jacques”, “Jean-Paul”, “Thomas” and “John” will respond correctly to “Nietzsche” with probability one-third simply by randomly guessing among the remaining responses, “Franz”, “Friedrich” and “Philip”. For this reason, observing a test taker incorrectly provides the response “John” to “Rousseau”, meaning the two responses are not independent.

In this paper, we develop an IRT framework for matching tests. The simplest model in this framework is the matching test Rasch model (MT-RM), but the framework is sufficiently general to incorporate more complex IRT models, like the 2PL. The matching test framework overcomes the difficulties inherent in modeling matching tests by developing a model of the response process. In particular, the matching test framework assumes that a test taker’s responses are determined by his or her knowledge of the test items. The probability that a test taker knows an item is determined by that item’s item response function (IRF). This approach was employed for dichotomous items in extending the 2PL to the three-parameter model (3PL) and 3PL to the four-parameter model (4PL) (respectively, Barton & Lord, 1981; Birnbaum, 1968).

Introducing a response process allows us to relate a test taker’s ability with his or her test responses, but it also introduces additional parameters indicating whether each test taker knows the correct response to each test item. These additional knowledge parameters make it difficult to perform inference in the matching test framework and make it difficult to compare the matching test model family with existing IRT methods. To deal with this, we show how the knowledge parameters can be marginalized from the matching test framework. This allows us to develop efficient inference algorithms. It also allows us to demonstrate an important prediction of the matching framework: ability-dependent guessing. This contrasts popular IRT accounts of guessing, like the 3PL, that predict a constant probability of guessing a correct response for all test takers. These differing model predictions motivate an empirical comparison between the MT-RM and an extension of the Rasch model accounting for guessing, the 1PL–G. This comparison shows that the MT-RM provides a better account of test-taking behavior, motivating its use in practical applications.

2. Matching IRT

2.1. Matching data

Consider a hypothetical test taker completing a test comprised of I matching items and D distractors. In the example in Fig. 1, the items are the philosophers’ last names in the left-hand column. To complete the test, the test taker must assign one of a set of I + D response alternatives to each of the I test items in such a way that no response alternative is assigned to more than one item. As previously noted, the D ≥ 0 distractors are not the correct response to any item. In Fig. 1, the response alternatives “Franz” and “Philip” are distractors.

One possible way of numbering the items in the sample test in Fig. 1 is shown in Table 1. In this table, the first two columns assign numerical labels to each of the five test items, and the last two columns assign numerical labels to each of the seven response alternatives. The items are numbered in the order that they appear in Fig. 1. The response alternatives are numbered in such a way that the first response alternative is the correct response to the first item, the second alternative is the correct response to the second item, etc.

<table>
<thead>
<tr>
<th>Number</th>
<th>Label</th>
<th>Response alternative</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Rousseau</td>
<td>Jean-Jacques</td>
</tr>
<tr>
<td>2</td>
<td>Sartre</td>
<td>Jean-Paul</td>
</tr>
<tr>
<td>3</td>
<td>Hobbs</td>
<td>Thomas</td>
</tr>
<tr>
<td>4</td>
<td>Nietzsche</td>
<td>Friedrich</td>
</tr>
<tr>
<td>5</td>
<td>Locke</td>
<td>John</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>Franz</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>Philip</td>
</tr>
</tbody>
</table>
Now suppose $P$ independent test takers complete the $I$ matching items. Each test taker $p = 1, \ldots, P$ offers a response $X_{pi}$ for each of the $I$ test items. Therefore, each row $X_{p}$ contains the responses of a single test taker to each of the $I$ items, and each column $X_{i}$ contains the responses of all $P$ test takers to a single item. Conceptually, each row of $X_{p}$ can be thought of as a sample of $I$ balls drawn without replacement from an urn containing $I + D$ balls, labeled 1, $\ldots$, $I + D$. The first ball drawn is the response to the first test item, the second ball is the response to the second, etc., resulting in a response pattern $X_{p}$ such that every response is an integer between one and $I + D$ and for any two items $i$ and $j$, $X_{pi} \neq X_{pj}$. In the special case where $D = 0$, this means that $X_{p}$ is a permutation of the first $I$ integers.

2.2. Knowledge and responding

To develop a model of matching test performance, we assume that every test taker completing a matching test either knows or does not know the correct response to each of the $I$ items and that his or her responses to the $I$ items are solely determined by this knowledge. Each test taker’s knowledge of the correct response to a test item is represented by an element of an $P \times I$ random matrix $K$, which we call the knowledge matrix. The elements of $K$ indicate whether a particular subject knows the correct response to a particular item: for any $p$ and $i$, $K_{pi} = 1$ if test taker $p$ knows the correct response to item $i$ and $K_{pi} = 0$ if $p$ does not know the correct response to $i$. The rows $K_{p}$ of $K$ indicate which of the $I$ items are known by test taker $p$, and the columns $K_{i}$ indicate which of the $P$ test takers know item $i$. Given test taker $p$’s knowledge $K_{p}$ of the $I$ items, $p$’s responses to the $I$ items are determined by the following cognitive process. First, the test taker assigns the correct response to every known item. Second, the test taker guesses the responses to the unknown items by randomly assigning one of the remaining responses to each of the remaining items.

Before developing the matching test framework in detail, it should be noted that, because the test taker always assigns the correct response to known items, the probability that a test taker incorrectly responds to a known item is zero. The same assumption is made by many other IRT models, including the Rasch, 2PL and 3PL. There is, however, currently some debate over when this assumption is justified. One the one hand, Barton and Lord (1981) found that allowing this probability to be greater than zero had very little effect on estimates of ability from three American academic exams, the Scholastic Aptitude Test, Graduate Record Exam and Advanced Placement exam. On the other hand, Osgood, McMorris, and Potenza (2002) and Reise and Waller (2003) determined that allowing this probability to be greater than zero was necessary to account for a self-reported delinquency scale and the Minnesota Multiphasic Personality Inventory (MMPI), respectively. Together these results suggest that our decision process is reasonable for the recognition tasks such as achievement tests, but may need to be generalized for other tasks.

2.2.1. From knowledge to response

The urn analogy from the previous section can help us to better understand the response process. Initially, the urn is populated with $I + D$ response alternatives. In the first step of the response process, the test taker correctly responds to each known item, meaning that he or she pulls each ball for which $K_{pi} = 1$ from the urn. Suppose that the test taker knows $S_{p} = \sum_{i=1}^{I} K_{pi}$ items, so that, after the first step, he or she has responded to $S_{p}$ items and those $S_{p}$ items have been removed from the urn. Then, in the second step, the test taker randomly draws $I - S_{p}$ response alternatives without replacement from the remaining alternatives in the urn. Since there are $I - S_{p} + D$ alternatives in the urn and each of these alternatives is selected at random, the probability a particular alternative is drawn for the first unknown item is $1/(I - S_{p} + D)$. Moreover, since alternatives are drawn without replacement, there will be one less alternative available as a response to the second unknown item than there was for the first, meaning that the probability the test taker selects any of the remaining responses $1/(I - S_{p} + D - 1)$.

Continuing this process until each of the $I - S_{p}$ unknown items have been answered, we find that the probability of any particular pattern of $I - S_{p}$ responses to the unknown items is

$$\frac{1}{I - S_{p} + D} \times \frac{1}{I - S_{p} + D - 1} \times \cdots \times 1 = \frac{D!}{(I - S_{p} + D)!}. $$

This allows us to specify $Pr(X_{p} \mid K_{p})$, the probability of an individual’s response pattern conditional on their knowledge of the items. Since the test taker always responds correctly to known items, the probability of a response pattern $X_{p}$ containing an item that is known, but incorrectly answered is zero. When all known items are correctly answered, the probability of a response $X_{p}$ is solely determined by the number of unknown items. Thus,

$$Pr(X_{p} \mid K_{p}) = \frac{D!}{(I - S_{p} + D)!} \text{ whenever } X_{pi} \text{ is correct for every item for which } K_{pi} = 1, \text{ and } Pr(X_{p} \mid K_{p}) = 0 \text{ otherwise. This latter probability is simply a statement that the probability of incorrectly responding to a known item is zero. Combining these probabilities across the } P \text{ independent test takers, we find}

$$Pr(X \mid K) = \prod_{p=1}^{P} Pr(X_{p} \mid K_{p}) = \prod_{p=1}^{P} \frac{D!}{(I - S_{p} + D)!} \text{ whenever } X_{pi} \text{ is correct for every test taker and item having } K_{pi} = 1, \text{ and } Pr(X \mid K) = 0 \text{ otherwise.}

2.2.2. From ability to knowledge

We assume that each test taker’s knowledge of a single item $i$ is determined by his or her ability $\theta$ by a continuous, strictly increasing surjection $F_{i} : \mathbb{R} \to (0, 1)$, known as $i$’s item response function (IRF). The assumption of surjectivity ensures the lower and upper asymptotes are respectively zero and one, as in the 2PL (Birnbaum, 1968). Additionally, we make the standard assumptions that $K_{p}$ is independent of $K_{p}$ for distinct test takers $p$ and $q$ and $K_{p}$ for distinct items $i$ and $j$ to obtain

$$Pr(K \mid \theta) = \prod_{p=1}^{P} \prod_{i=1}^{I} F_{i}(\theta_{p}),

where $\theta = (\theta_{1}, \ldots, \theta_{P})$. Combining this expression and Eq. (2) defines the matching test model.

We can obtain a number of important special cases by restricting $F_{i}(\theta) = F(\alpha_{i}(\theta - \beta))$, where $F$ is either the logistic (2PL) or normal (2PNO) cumulative density function (CDF). We will call the former model the MT-2PL and the latter the MT-2PNO. By imposing the additional constraint that $\alpha_{i} = 1$, one obtains matching test versions of the Rasch (MT-RM) and 1PNO (MT-1PNO) models.

2.2.3. Knowledge marginalization

For most psychometric applications, we are not interested in estimating $K$. In this case, we can marginalize $K$ from the likelihood. Since the marginalized model only contains continuous variables, it can be efficiently sampled using Hamiltonian Monte Carlo (Gelman et al., 2014). Marginalization is likely also beneficial for maximum likelihood estimators, since finite random variables are also problematic in these contexts.
We begin by supposing that $X_{pi}$ is correct for every item $i$. In this case, $Pr(X_p | K_p) > 0$ every knowledge pattern $K_p$. We can compute the marginal probability $Pr(X_p | \theta_p)$ by marginalizing the joint probability $Pr(X_p, K_p | \theta_p)$ with respect to $K_p$. Define $K = \{0, 1\}$ to be the set of $i$-element binary vectors. Then,

$$Pr(X_p | \theta_p) = \sum_{K_p \in K} Pr(X_p, K_p | \theta_p) = \sum_{K_p \in K} Pr(X_p | K_p) \cdot Pr(K_p | \theta_p)$$

$$= \sum_{K_p \in K} \left[ \frac{D!}{(I - S_p + D)!} \right] \cdot Pr(K_p | \theta_p)$$

$$= \sum_{i=0}^I \left[ \frac{D!}{(I - S_p + D)!} \right] \cdot Pr(S_p = s | \theta_p),$$

since $D!/(I - S_p + D)!$ only depends on $K_p$ through $S_p$. We can obtain a more compact expression by noting that the final expression is equal to the expected value of $D!/(I - S_p + D)!$ over the distribution of $Pr(S_p | \theta_p)$, i.e.,

$$Pr(X_p | \theta_p) = \mathbb{E}_{S_p | \theta_p} \left[ \frac{D!}{(I - S_p + D)!} \right] \quad (4)$$

The probability $Pr(S_p = s | \theta_p)$ is the probability that $p$ knows the correct responses to $s$ items. It is the sum of the probabilities of all of the response patterns $K_p$ with $s$ ones and can be computed for any $\theta_p$ and IRF using the Lord–Wingersky algorithm (Lord & Wingersky, 1984), a recursive algorithm for computing the sum score distribution of an IRT model without having to enumerate every possible response pattern. Define $S_p^{1:s} = \sum_{i=0}^s K_p^i$. The insight underlying the Lord–Wingersky algorithm is to notice that $S_p^{1:s} = K_p 1$ and $S_p^{1:s} = S^{1:(i-1)} + K_p^i$. This allows us to compute the distribution of $S_p^{1:s}$ by convolving $S_p^{1:(i-1)}$ with $K_p^i$. Since $S_p = S_p^{1:s}$, we can compute $S_p$ from the distribution of $K_p^i$ and then applying the recursive relationship between the successive $S_p^{1:s}$ starting with $i = 1$ to obtain $S_p$. We can then use these probabilities to compute the marginal likelihood $Pr(X_p | \theta_p)$ using Eq. (4).

We can use the same technique to marginalize $K_p$ when some $X_{pi}$ are incorrect. Define $V_{pi}$ to be the set of items incorrectly answered by $p$. The matching test model assumes that every known item is answered correctly. As a consequence, $Pr(X_p | K_p) = 0$ for every knowledge pattern having $K_p = 1$ for at least one item $i \in V_{pi}$. Define $R_p$ to be the number of correct items. For the $R_p$ correct items, we can apply the previous argument, replacing $I$ with $R_p$. The remaining items will have $K_p = 0$ for every knowledge pattern with positive probability. Thus, the marginal likelihood is

$$Pr(X_p | \theta_p) = \left[ \prod_{i \in V_p} Pr(K_p = 0 | \theta_p) \right] \cdot \mathbb{E}_{S_p | \theta_p} \left[ \frac{D!}{(K_p - S_p + D)!} \right] \quad (5)$$

The expectation in this expression can be computed by applying the Lord–Wingersky algorithm to the set of correct items, $V_p$.

### 2.3. Extension to multiple blocks

Suppose we would like to add another block of items to our matching test, whose items are independent from those in Fig. 1. For example, we might add a new block whose items are “Kant”, “Fodor”, “Leibniz”, “Russell” and “Quine” and whose response alternatives are “Immanuel”, “Jeffrey”, “Gottfried”, “Bertrand” and “Willard”. The first five items of the resulting test are the items from the first block and the second five are the items from the second. In general, for a test containing $B$ blocks, the first $I_1$ items belong to the first block, the second $I_2$ to the second block, etc. Define $I_0 = 0$ and $I_b^* = \sum_{j=1}^{b} I_j$. Then, the items belonging to set $b$ are $I_{b-1}^* + 1, \ldots, I_b^*$.

Item dependence may arise from two sources in tests comprised of multiple blocks. First, dependence may arise from the matching structure of each block. We capture this dependence by modeling each block’s responses using the likelihood in Eq. (2) for each block. To this end, suppose block $b$ has $I_b$ items and $D_b$ distractors. Moreover, define $X_p^{(b)}$ and $K_p^{(b)}$ to be the responses and knowledge of test taker $p$ for the items of block $b$, and suppose that test taker $p$ knows $S_{pb}$ items in block $b$. Then, the likelihood of $X$ given $K$ is

$$Pr(X | K) = \prod_{b=1}^B \prod_{p=1}^{\sigma_b} \frac{D_b!}{(I_b - S_{pb} + D_b)!}$$

whenever $X_{pb}$ is correct whenever $K_{pi} = 1$. Otherwise, the likelihood is zero.

The second source of dependence is the correlational structure in the item knowledge within each block. We deal with this source of dependence using a testlet model (Bradlow, Wainer, & Wang, 1999). Testlet models were developed to account for test-taker behavior on groups of items related through a common setup. For example, suppose a test contains four items testing comprehension of a common text passage. Intuitively, we expect a test taker’s responses to the items to be related not only through the test taker’s ability, but also through their understanding of the passage, violating the local independence assumption.

Testlet models allow a test taker’s ability to vary over blocks. This allows them to account for the additional correlation between items within a block, or testlet, using standard IRT models. To do this, testlet models decompose the ability of test taker $p$ on the items of block $b$ into $\theta_p + \eta_{pb}$, where the $\eta_{pb}$ are independent, identically-distributed samples from normal distribution with mean zero and variance $\sigma_{\eta}^2$. Thus, the expected ability of test taker $p$ in block $b$ is simply his or her ability $\theta_p$ and the expected deviation across test takers and blocks is $\sigma_{\eta}$.

To account for correlations in item knowledge, we apply the testlet model to the knowledge parameters $K_{pi}$. Suppose the test items are split among $B$ blocks. The probability that test taker $p$ knows item $i$ in block $b$ is $F(\theta_p + \eta_{pb})$. Now define $\eta$ to be the $P \times B$ matrix whose elements are $\eta_{pb}$, and let $i \in b$ denote the situation where item $i$ is from block $b$. Then, the probability of a knowledge matrix $K$ is

$$Pr(K | \theta, \eta) = \prod_{p=1}^P \prod_{b=1}^B \prod_{i \in b} F(\theta_p + \eta_{pb})$$

$$\eta_{pb} \sim N(0, \sigma_{\eta}^2)$$

The full matching test model for $B$ blocks can be obtained by combining Eq. (6) with Eq. (7).

We could use the same approach to incorporate blocks of matching items into longer tests. For example, to incorporate matching blocks into a test otherwise comprised of multiple choice items, we could apply the 2PL or 3PL model to the multiple choice items and apply the matching model, with testlets, to the matching items. The likelihood of a test taker’s responses is the product of the likelihood of their responses to the multiple choice items and the likelihood of their responses to the matching items under the testlet matching test model just presented.
3. Properties of the matching test framework

In this section, we derive the marginal item response function and the distribution of sum scores for matching test models. For simplicity, we focus on tests comprised of a single block, since their generalization to tests comprised of multiple blocks is obvious. The properties developed here will later be used to demonstrate the practical utility of the matching test framework.

3.1. Marginal item response function

Recall that the IRF of an MT gives the probability of correctly answering to an item as a function of $\theta$. This function is useful in comparing the matching model to existing IRT models. In this section, we derive the IRF of the matching model and investigate some of its properties. In general, an IRF is a continuous, strictly increasing function $G_i$ from $\mathbb{R}$ onto an open interval of $(0, 1)$. We will later prove that for the matching model, this interval is $(\frac{1}{D}, 1)$. In deriving the IRF, it is useful to suppress the person subscript $p$ and define random variables $K_i$ and $s_i$, denoting the knowledge and response, respectively, of a test taker of arbitrary ability $\theta$. Further, we define a random variable $Y_i$, which is one whenever $K_i$ is correct and zero otherwise.

To compute $G_i$, we first note that $G_i(\theta) = \Pr(Y_i = 1 \mid \theta)$. We can expand $\Pr(Y_i = 1 \mid \theta)$ in $K_i$ to obtain $G_i(\theta) = \Pr(Y_i = 1 \mid K_i = 1, \theta) \cdot \Pr(K_i = 1 \mid \theta) + \Pr(Y_i = 1 \mid K_i = 0, \theta) \cdot \Pr(K_i = 0 \mid \theta)$. Now define $\gamma_i(\theta) = \Pr(Y_i = 1 \mid K_i = 0, \theta)$. Then, $G_i(\theta) = \Pr(Y_i = 1 \mid K_i = 1, \theta) \cdot F_i(\theta) + \gamma_i(\theta) \cdot [1 - F_i(\theta)]$.

since the matching model assumes that known items are answered correctly, meaning that $\Pr(Y_i = 1 \mid K_i = 1) = 1$. Rearranging terms, we obtain

$$G_i(\theta) = \gamma_i(\theta) + [1 - \gamma_i(\theta)] \cdot F_i(\theta). \tag{8}$$

Finally, we compute the probability of guessing the correct answer to the item, $\Pr(Y_i = 1 \mid K_i = 0, \theta)$. To this end, it is instructive to recall the ways that an unknown item can be correctly answered in our matching framework. Suppose that $K_i = 1$ for every item other than $i$. When there are no distractors, the probability of a correct response is one, since the only available response to $i$ is the correct one. In the more general case with $D \geq 0$ distractors, the test taker selects at random between one correct response and $D$ distractors, so the probability of a correct response is $1/(1 + D)$.

Now suppose that $K_i = 0$ and $K_i = 0$ for one other item, $i'$. When there are no distractors, there are two possible response patterns, one which correctly answers $i$ and $i'$, and another which incorrectly swaps the correct answers $i$ and $i'$. When there are $D$ distractors, there are $D!/(2 + D)!$ possible responses, of which $D!/(1 + D)!$ offer the correct response to $i$. Thus, the probability of a correct response is

$$D!/(2 + D)! = \frac{1}{2 + D}.$$

More generally, suppose that $K_i = 0$, but $K_i = 1$ for $S - i$ of the remaining test items. This means that $K_i = 0$ for $I - S - i$ items besides $i$. For example, when every item except $i$ is known, $S - i = I - 1$, and when one item besides $i$ is unknown, $S - i = I - 2$. The previous examples highlight the fact that, when $K_i = 0$, the probability of correctly responding to $i$ depends on the number $S - i$ of the remaining test items that are known. This means that

$$\gamma_i(\theta) = \sum_{s=0}^{I-1} \Pr(Y_i = 1 \mid K_i = 0, S - i = s) \cdot \Pr(S - i = s \mid \theta). \tag{9}$$

We are familiar with $\Pr(S - i = s \mid \theta)$, since this distribution arose in the context of knowledge marginalization. It can be computed for any value of $\theta$ using the Wingersky–Lord algorithm. We have already computed $\Pr(Y_i = 1 \mid K_i = 0, S - i = I - 1) = 1/(1 + D)$ and $\Pr(Y_i = 1 \mid K_i = 0, S - i = I - 2) = 1/(2 + D)$. In the general case where $S - i = s$, we have $D!/(I - s + D)!$ possible responses to the $i - s$ unknown items, of which $D!/(I - s + D - 1)!$ offer the correct response to item $i$. Thus,

$$\Pr(Y_i = 1 \mid K_i = 0, S - i = s) = \frac{1}{I - s + D}.$$

We can substitute this into Eq. (9) to obtain

$$\gamma_i(\theta) = \sum_{s=0}^{I-1} \frac{1}{I - s + D} \cdot \Pr(S - i = s \mid \theta)$$

$$= \mathbb{E}_{s \sim p} \left( \frac{1}{I - s + D} \right). \tag{10}$$

We can use Eqs. (5) and (10) to compare the MT-2PL with the 3PL model. The IRF of the 3PL model can be written as

$$\gamma_i + (1 - \gamma_i) \cdot F_i(\theta),$$

where $F_i(\theta)$ is the IRF of the 2PL model. In the 3PL model, the guessing parameter $\gamma_i$ is constant for each item $i$. In the matching test model, the guessing parameter $\gamma_i(\theta)$ depends on the ability $\theta$ of the test taker.

The dependence of the guessing parameter on ability is demonstrated by Fig. 2, which plots the guessing parameter $\gamma_i(\theta)$ for the MT-2PL for different numbers of items, distractors and item difficulties. For simplicity, the discrimination parameter $\alpha_i$ was set to one for all items. For each plot, we generated a test of $I$ items, each having difficulty $\beta_i \in (-1, 0, 1)$. We then computed $\gamma_i(\theta)$ for each $\theta$ in an evenly-spaced grid of 101 values ranging from $-5$ to $5$ for each $\beta_i, D$ and $I$. These plots show that $\gamma_i(\theta)$ is an increasing $\theta$ and decreasing in $\beta$. Additionally, they show that the lower asymptote of $\gamma_i(\theta)$ decreases in both $I$ and $D$, while the upper asymptote only decreases in $D$. These properties are stated formally in the following theorem, which is proved in Appendix.

Theorem 1. For a given item $i$, $\gamma_i$ is a continuous, strictly increasing surjection from $\mathbb{R}$ and $(\frac{1}{1+D}, 1)$. Moreover, in the special case where $F_i(\theta) = F(\alpha_i(\theta - \beta_i))$ for all $i \in \{1, \ldots, I\}$, $\gamma_i$ is strictly decreasing in $\beta_i$ for each $j \neq i$.

A consequence of Theorem 1 is that the function $G_i$ defined by Eq. (8) is an IRF whose range is $(\frac{1}{1+D}, 1)$ as long as $F_i$ is an IRF. This result is stated in the following corollary, which is also proved in the Appendix.

Corollary 1. For any IRF $F_i$, the function $G_i$ defined by Eq. (8) is an IRF whose range is $(\frac{1}{1+D}, 1)$.

3.2. Score distribution

In this section, we derive the distribution over scores $T_p = \sum_{i=1}^{m} Y_p$ for a test taker whose ability is $\theta_p$. Later, we will use this distribution to compare the MT-2PL to the 3PL model. An expression for the distribution of $T_p$ is given by the following theorem, which is proved in the Appendix.

Theorem 2. Let

$$\pi_{m}^D = 1 - \sum_{u=1}^{m} (-1)^{u+1} \cdot \binom{m}{u} \cdot \frac{(m + d - w)!}{(m + d)!}.$$

Then,

$$\Pr(T_p = k) = \binom{I - S_p}{T_p - S_p} \cdot \frac{(I - T_p + D)!}{(I - S_p + D)!} \cdot \pi_{m}^D.$$


Distribution of depends on the number of distractors

\[ \text{Corollary 2.} \quad T \text{ stochastically dominates } S \]

\[ \text{Fig. 2. } \gamma_\theta \text{ as a function of ability for tests of length 5 or 10 with either 0 or 2 distractors.} \]

and

\[ \Pr(T_p | \theta_p) = \sum_{s \leq T_p} Pr(T_p | S_p = s) \Pr(S_p = s | \theta_p). \]

Despite its complicated statement, Theorem 2 is largely intuitive. The first term, \( \pi_{d}^{m} \), is the probability of incorrectly answering every item of an \( m \) item test with \( d \) distractors. The second, \( \Pr(T_p | S_p) \), combines the probability of the incorrect items, \( \pi_{d}^{m} \), with that of the correct items for a given number of known items. We average \( \Pr(T_p | S_p) \) over every number known less than or equal to the number of items correct.

Fig. 3 illustrates how the distribution of \( T_p | \theta_p \) (Correct) compares to distribution of \( S_p | \theta_p \) (Known). Each panel plots \( \Pr(T_p \geq \tau | \theta_p) \) and \( \Pr(S_p \geq \tau | \theta_p) \) – respectively, the probabilities of correctly answering and knowing at least \( \tau \) items – for every possible value of \( \tau \) for a 10 item test. It shows that \( \Pr(T_p \geq \tau | \theta_p) \geq \Pr(S_p \geq \tau | \theta_p) \), consistent with the findings of previous authors (e.g., Benson & Crocker, 1979; Shaha, 1984) that test takers tend to score higher on matching tests than on tests comprised of dichotomous items. In fact, we show in Corollary 2 (below) that \( \Pr(T_p \geq \tau | \theta_p) > \Pr(S_p \geq \tau | \theta_p) \), i.e., that \( T_p | \theta_p \) stochastically dominates \( S_p | \theta_p \).

\[ \text{Corollary 2.} \quad T_p \text{ stochastically dominates } S_p. \]

Fig. 3 also demonstrates how the distribution of \( T_p | \theta_p \) depends on the number of distractors \( D \). It shows that the distribution of \( T_p \) looks more like the distribution of \( S_p | \theta_p \) when \( D = 2 \) than when \( D = 0 \), as suggested by Theorem 1. Perhaps more interestingly, the left panel shows that the probability of correctly answering nine items is equal to the probability of correctly answering ten items for a test with \( I = 10 \) and \( D = 0 \). The reason is intuitive: when there are no distractors, there will only be one response left for the last item.

4. Empirical evaluation

4.1. Data and design

In this section, we empirically assess whether the matching test framework provides a better account of human behavior than existing IRT models using unpublished data collected by Napolitano, Batchelder and Steyvers. This dataset contains responses from 114 undergraduate students in an undergraduate psychology course at the University of California, Irvine to six matching tests comprising different areas of common knowledge. The areas tested were Greek letters, philosophers, flags, United States presidents, car logos and translations of the phrase “Happy New Year”. The instructions shown to test takers are shown in Table 2.

To evaluate the matching test framework, we need to select a benchmark IRT model. The 3PL is a natural choice. The previous section demonstrates that a key difference between matching test models and standard models for dichotomous items is their account of guessing behavior. As a consequence, it is important to use a benchmark that can account for guessing. Otherwise, it would be unclear whether differences in model performance stem from the specifics of the matching test framework’s account of guessing or the fact that it provided an account of guessing at all.

Because there were only a small number of test takers in the dataset, we fit the data with the MT-RM. Additionally, we restricted the discrimination parameter in the 3PL to be one. This model is often called the one-parameter logistic model with guessing, or 1PL-G. Employing the Rasch model in both comparisons means that the only difference between the two models is their account of guessing. The MT-RM offers an ability-dependent account of guessing that is specific to matching tests, while the 1PL-G assumes that there is a constant guessing rate for each item.

4.2. Evaluation criterion

We compare the two models on their ability to predict test taker responses to the items of the matching tests. Let \( n_i \) be the
number of test takers answering t test items correctly. We can summarize the responses of the test takers using the vector \( n = (n_0, n_1, \ldots, n_I) \), typically called the observed score distribution. Any IRT model predicts the score distribution \( T_p \) for each test taker \( p \) given its model parameters \( \psi \). For the matching model, \( \psi = (\theta, \beta) \). The expression for \( T_p \) is given by Theorem 2. For the 1PL-G, \( \psi = (\theta, \beta, \gamma) \). \( T_p \) can be computed by applying the Lord–Wingersky algorithm.

These distributions can be combined to define a probability distribution over the observed scores \( n_t \), which we denote by the random vector \( N = (N_0, N_1, \ldots, N_I) \). The distribution of each \( N_t \) is the sum over all \( N_t \) element subsets of the \( P \) test takers of the probability that all \( N_t \) test takers correctly answer \( t \) test items, \( \sum P(T_p = t | \psi) \).

Béguin and Glas (2001) suggest evaluating model fit by comparing the observed score distribution \( n \) to its expectation under an IRT model using the \( \chi^2 \) discrepancy statistic,

\[
\chi^2 = \sum_{t=0}^{I} \frac{[N_t - E(N_t | \psi)]^2}{E(N_t | \psi)}.
\]

In a frequentist framework, we could compute a point estimate of the IRT model parameters \( \hat{\psi} \) using a joint, marginal or conditional maximum likelihood inference approach and plug this estimate into \( \chi^2 \) in the above expression. Unfortunately, the sampling distribution of \( \chi^2 \) does not follow a \( \chi^2 \) distribution, so we would have no way of interpreting and comparing the values of \( \chi^2 \) between the models.

We employ a Bayesian approach suggested by Sinharay, Johnson, and Stern (2006) to remedy for this situation. Bayesian inference for IRT models is based on the posterior density of the model parameters

\[
p(\psi | X) = \frac{Pr(X|\psi)p(\psi)}{\int_\psi Pr(X|\psi)p(\psi) d\psi}.
\]

The first term in the numerator of the righthand side of this expression, \( Pr(X|\psi) \), is the likelihood of the IRT model. For the matching model, this is given by Eq. (5). For the 1PL-G, it is

\[
Pr(X|\psi) = \prod_{p=1}^{P} \prod_{i=1}^{I} \left[ \gamma_i + (1 - \gamma_i) \cdot \frac{e^{\psi_i(n_p - \beta_i)}}{1 + e^{\psi_i(n_p - \beta_i)}} \right].
\]

The second, \( p(\psi) \), is the prior distribution. We use a standard prior distribution in Bayesian IRT modeling wherein all of the model parameters are independent. To ensure that the parameters are identified, the prior distribution of the ability parameters \( \theta_p \) is taken to be \( N(0, 1) \). The prior distribution of the difficulty parameters \( \beta_i \) is taken to be \( N(\mu_\beta, \sigma_\beta^2) \), where \( \mu_\beta \) has an improper uniform prior and \( \sigma_\beta^2 \) has an inverse-\( \chi^2 \) distribution with both scale parameters equal to 1/2. In the 1PL-G, the prior distribution of the \( \gamma_i \) is taken to be a uniform distribution on the unit interval.

Instead of plugging in a point estimate for \( \psi \), Sinharay et al. (2006) suggest taking the expectation of \( N_t \) over the posterior density of the model parameters, \( p(\psi | X) \). Denote this value

\[
\chi^2_{\text{obs}} = \sum_{t=0}^{I} \frac{[n_t - E_{\psi | X}(N_t)]^2}{E_{\psi | X}(N_t)}.
\]  

We can compare \( \chi^2_{\text{obs}} \) to its Bayesian posterior predictive distribution. The posterior predictive distribution \( Pr(\hat{X} | X) \) is the predictive distribution of new response matrices \( X \) over the posterior distribution, i.e., \( \int_\psi Pr(X | \psi)p(\psi | X) d\psi \). We can use \( Pr(\hat{X} | X) \) to compute the posterior predictive distribution of the log discrepancy ratio \( \log(\chi^2_{\text{obs}}/\hat{\chi}^2_{\text{m}}) \) for every response matrix \( X \) and marginalizing.

These quantities are easy to compute using Markov chain Monte Carlo (MCMC) sampling. MCMC methods are popular for Bayesian inference, because they can draw samples from \( p(\psi | X) \) without computing the normalizing factor \( \int_\psi Pr(X | \psi)p(\psi) d\psi \), which is intractable in most applications. We can use MCMC methods to sample the posterior predictive distribution of the log discrepancy ratio as follows. First, we drew 1000 samples \( \hat{\psi}^{(m)} \) from \( p(\psi | X) \) after a 1000 sample burn-in on each of four independent MCMC chains seeded from random samples from the joint prior distribution. We assessed convergence using the split-chain \( \hat{R} \) (Gelman et al., 2014). This statistic was below 1.1 for each parameter, consistent with convergence to the joint posterior. We provide code for efficiently sampling the two models in Stan (Zeigenfuse & Steyvers, 2020).

For each of these samples, we sampled a response matrix \( \hat{X}^{(m)} \) from \( Pr(\hat{X} | X) \) and compute its observed score distribution \( n^{(m)} \). Given the collection of \( M \) samples from the posterior, we can estimate \( E_{\psi | X}(N_t) \) using the Monte Carlo estimator

\[
E_{\psi | X}(N_t) = \sum_{m=1}^{M} \sum_{p=1}^{P} Pr(T_p = t | \psi^{(m)}).
\]  

We then compute the Monte Carlo estimate for \( \chi^2_{\text{obs}} = \hat{\chi}^2_{\text{obs}} \) by substituting \( E_{\psi | X}(N_t) \) for \( E_{\psi | X}(N_t) \) in Eq. (11). We also compute

\[
\hat{\chi}^2_{\text{m}} = \sum_{t=0}^{I} \frac{[n_t^{(m)} - E_{\psi | X}(N_t)]^2}{E_{\psi | X}(N_t)}.
\]

Finally, we compute the log discrepancy ratio \( \log(\chi^2_{\text{obs}}/\hat{\chi}^2_{\text{m}}) \) for each posterior sample \( \psi^{(m)} \).
**Table 3** Posterior predictive $p$-value for each model and data set.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>MT-RM</th>
<th>1PL-G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Balls</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Car Logos</td>
<td>0.158</td>
<td>0.0085</td>
</tr>
<tr>
<td>Flags</td>
<td>0.06</td>
<td>0.0075</td>
</tr>
<tr>
<td>Greek Letters</td>
<td>0.179</td>
<td>0.0005</td>
</tr>
<tr>
<td>Happy New Year</td>
<td>0.0045</td>
<td>0.0</td>
</tr>
<tr>
<td>Philosophers</td>
<td>0.2875</td>
<td>0.003</td>
</tr>
<tr>
<td>Presidents</td>
<td>0.0075</td>
<td>0.0</td>
</tr>
</tbody>
</table>

4.3. Fit comparison

Fig. 4 shows the log discrepancy ratio $\log(\chi^2_{obs}/\chi^2_{mod})$ for each posterior sample, model and dataset. These ratios indicate how well each model fits each data set. Positive values indicate that the observed fit is worse than the fit to the $m$th posterior predictive sample. The farther above zero, the worse the fit. The area each violin has below zero is an estimate of the probability of the observed sample under the posterior predictive distribution, its posterior predictive $p$-value. These values are shown in Table 3.

Fig. 4 and Table 3 demonstrate that overall the observed data is more consistent with the MT-RM model. For every data set, the bulk of the distribution of sampled log discrepancy ratios is closer to zero for the MT-RM than the 1PL-G. This translates into lower posterior predictive $p$-values for the 1PL-G for every data set, indicating that the observed score distribution is more likely under the MT-RM than under the 1PL-G. Closer examination of the $p$-values reveals that every observed score distribution is highly unlikely under the 1PL-G. By contrast the observed data is reasonably likely under the MT-RM for the car logos, Greek letters, flags and philosophers data sets. The poor fit to the remaining data sets is likely attributable to the small sample size.

4.4. Parameter estimates

Fig. 5 shows the distribution of widths – which we define to be the difference between the largest and smallest sample – of the posterior distributions of the guessing parameters in the 1PL-G for each data set. These widths provide an index of the uncertainty about the probability of correctly guessing the response to each test item. Inspection of Fig. 5 shows that there is a considerable amount of variability in the posterior distribution of each guessing parameter. **Theorem 1** leads us to expect wider posterior distributions for the guessing parameter when a matching test model holds, since the matching test model predicts that the effective guessing parameter will vary considerably across test takers. This should result in posterior uncertainty in a model, like the 1PL-G, that does not allow variability.

5. Discussion

In this paper, we presented an IRT framework for estimating latent traits and abilities using matching tests. Our framework offers a model of how matching test responses arise from item knowledge and use IRT models for dichotomous items to model knowledge. Particular choices of IRT models generate different matching test models. For example, the MT-RM arises when we use a Rasch model to model knowledge. We also demonstrated how to marginalize the knowledge parameter and derived a number of important properties of the matching test framework, including its item response function and sum score distribution. We saw that the matching test framework predicts ability-dependent guessing, motivating an empirical comparison between the MT-RM and the 1PL-G. This comparison demonstrated that the MT-RM provided a better account of test-taking behavior than the 1PL-G, suggesting that the matching test framework is important in practical applications.

These results suggest that matching tests may be a viable alternative to multiple-choice tests for educational and psychometric measurement, when the matching test framework is used for trait and ability estimation. Without the matching test framework, however, considerable care must be taken when applying matching tests. As demonstrated by the model's derivation and our empirical study, matching test takers have a higher probability of correctly answering an item for a fixed ability and difficulty than multiple-choice test takers. Thus, raw scores on these tests should not be compared without accounting for the difference in these probabilities. Moreover, it can likely explain the differences in test performance found in studies comparing matching an multiple-choice items (Benson, 1981; Benson & Crocker, 1979; Haynie, 2003; Shaha, 1984).

Though the current paper provides a solid foundation for matching tests, more work is still needed. One drawback of the current framework is its inability to capture partial knowledge. In the philosopher's example, for instance, the test taker might know the correct responses to "Locke" and "Hobbes", but also that Nietzsche was German. The test taker would know that "John" and "Thomas" were not the correct answers for "Nietzsche", since they were the correct answers to "Locke" and "Hobbes". Moreover, they could use their knowledge that Nietzsche is German to guess that "Jean-Jacques" and "Jean-Paul" are not correct answers to "Nietzsche". This would allow the test taker to eliminate these two response alternatives, improving their chances of a correct response from one in five to one in three.

Accounting for partial knowledge on matching tests is challenging. In the previous example, the test taker used knowledge of the item and response alternatives in order to rule out possible response alternatives. This suggests that adequately accounting for partial knowledge will require an interaction between the test taker, test item and response alternative. One possibility in this vein would be to employ an ordinal IRT model for the knowledge parameter, such as the rating scale or partial credit model. Ordinal IRT models are used in situations where we have three or more ordered categories of responses. Common examples of such situations are rating scales (e.g., "Disagree", "Neutral", "Agree") and partial credit on achievement tests.

To apply an ordinal model to the matching framework, we could define each knowledge parameter $K_i$ to have three ordered states: no knowledge, partial knowledge and full knowledge. The "no knowledge" and "full knowledge" cases are analogous to the situations where $K_{pi} = 0$ and $K_{pi} = 1$ in our model. In the "partial knowledge" case, the test taker is not able to produce the correct answer, but is able to exclude some response alternatives.

The challenge in this approach is determining which response alternatives can be excluded on the basis of partial knowledge. Ideally, we would like to learn from test data which response alternatives are excluded by test takers with partial knowledge. To this end, we could define a matrix whose elements give the probability that a given response alternative would be offered as a response to a given test item when a test taker has partial knowledge. Parameter inference is the key challenge for such a model. It will likely require large amounts of data and, potential, parameter restrictions. That said, we could likely apply the same technique for marginalizing the knowledge parameter in order to simplify inference.

A related way of dealing with partial knowledge is to group matching items into blocks of closely related items, mitigating its effects. In our example of partial knowledge, the test taker was able to leverage name origins to improve their chances of a correct response. If instead we divided French, German and
English philosophers into separate blocks of matching items, we could limit the test taker’s capacity to apply this type of partial knowledge.

In addition to dealing with partial knowledge, future work should evaluate the performance of the MT-2PL model, both when there are multiple blocks of matching items and when matching items have been incorporated into larger tests of dichotomous items. We developed extensions for these situations in the section extending the matching test framework to multiple blocks. Understanding the behavior of the matching model in these situations is important for understanding when matching items can be used for psychometric measurement.

Finally, methods for detecting differential item functioning (DIF) should be developed. Detecting DIF is important for ensuring fair tests. For example, suppose that an item is easier for women than it is for men. Then, we expect women to correctly answer the item more often than men, potentially leading to higher test scores and higher estimated ability. The standard Rasch model and other IRT models are often used to detect DIF (e.g., Thissen, Steinberg, & Wainer, 2012), since it can be measured in group difference in item parameters. Thus, by developing methods for detecting group differences in the parameters of the matching test models, we can detect DIF in matching tests. By implementing this and the previous methods, practitioners who desire to use matching tests can enjoy the benefits that IRT methods proffer to dichotomous and polytomous test items.

Appendix

We make use of the following two lemmas to prove Theorem 1.

Lemma 1. Define $S^{1f} = \sum_{i=1}^{I} k_i$, and $H_{1,s}(\theta) = \Pr(S^{1f} \geq s \mid \theta)$. For all $s \geq 1$, $H_{1,s}$ is a continuous, strictly increasing function with 
\[
\lim_{\theta \to -\infty} H_{1,s} = 0 \quad \text{and} \quad \lim_{\theta \to \infty} H_{1,s} = 1.
\]
Proof. By induction on \( l \). When \( l = 1 \), \( H_{1,1}(\theta) = F_1(\theta) \), which is continuous and has the desired limits by definition. Suppose the lemma holds for \( I = J \). We can express \( H_{I,1+s}(\theta) \) in terms of \( F_{I,1} \), \( H_{I,1} \) and \( H_s \) as
\[
H_{I,1+s}(\theta) = F_{I,1}(\theta) \cdot H_{I,1-s}(\theta) + [1 - F_{I,1}(\theta)] \cdot H_s(\theta),
\]
which is clearly continuous. Choose \( \theta' < \theta \). Then,
\[
H_{I,1+s}(\theta') - H_{I,1+s}(\theta)^{'} > F_{I,1}(\theta') \cdot [H_{I,1}(\theta') - H_{I,1}(\theta')]
\]
since \( F_{I,1} \) is strictly increasing. Moreover, by the inductive assumption we have \( H_{I,1-s}(\theta') - H_{I,1-s}(\theta') \) and \( H_s(\theta') - H_s(\theta') \).

Finally, note that
\[
\lim_{\theta \to -\infty} H_{I,1+s}(\theta) = H_{I,1+s}(\theta) = 0
\]
\[
\lim_{\theta \to +\infty} H_{I,1+s}(\theta) = H_{I,1-s}(\theta) = 1
\]
by combining the definition of \( F_{I,1} \) with the inductive hypothesis. □

Lemma 2. Suppose \( F_2(\theta) = F(\alpha_1(\theta - \beta_1)) \) for all \( i \in \{1, \ldots, I\} \), and define \( \alpha = (\alpha_1, \ldots, \alpha_I) \) and \( \beta = (\beta_1, \ldots, \beta_I) \). Additionally, for fixed \( \alpha \) and \( \beta \) define \( H_{I,1,s}(\beta) = Pr(S \geq s \mid \theta, \alpha, \beta) \). Then, \( H_{I,1,s} \) is continuous and strictly decreasing in each \( \beta_i \).

Proof. By induction on \( I \). The base case follows from the argument of Lemma 1. Suppose the lemma holds for \( I = J \), and define \( \beta_{I,J} = (\beta_1, \ldots, \beta_J) \). Then,
\[
H_{I,1+s,1}(\beta) = F(\alpha_I(\theta - \beta_I)) \cdot H_{I,1-s,J}(\beta_1)
\]
\[
+ [1 - F(\alpha_I(\theta - \beta_I))] \cdot H_s(\beta_1).
\]
Noting that \( F(\alpha_I(\theta - \beta_I)) \) is continuous in \( \beta_I \) and applying the inductive hypothesis, we obtain that \( H_{I,1,s} \) is continuous all \( \beta_i \). Select \( \beta^*_{I,J} < \beta_{I,J} \), and define \( \beta^* = (\beta_{I,J}, \beta^*) \). Then,
\[
H_{I,1+s,J}(\beta^*) - H_{I,1+s,J}(\beta) = F(\alpha_I(\theta - \beta_I)) \cdot [H_{I,1-s,J}(\beta_1) - H_{I,1-s,J}(\beta_1)]
\]
\[
+ [1 - F(\alpha_I(\theta - \beta_I))] \cdot [H_s(\beta_1) - H_s(\beta_1)] = 0.
\]
The remainder of the lemma is obtained from the inductive hypothesis. □

Proof of Theorem 1. Without loss of generality, we prove the theorem for the first item. For any \( I \geq 2 \),
\[
\gamma_I(\theta) = \frac{1}{1 + \sum_{s=I}^{I+D} \left( \frac{1}{I - s + D} - \frac{1}{I - s + 1 + D} \right) \cdot H_{I-1,s}(\theta)}.
\]
By Lemma 1, \( \gamma_I \) is continuous. Now select \( \theta' < \theta \). Then,
\[
\gamma_I(\theta) - \gamma_I(\theta') = \sum_{s=I}^{I+D} \left( \frac{1}{I - s + D} - \frac{1}{I - s + 1 + D} \right) \cdot [H_{I-1,s}(\theta) - H_{I-1,s}(\theta')]
\]
also by Lemma 1. Additionally,
\[
\int_{\theta \in \mathbb{R}} \gamma_I(\theta) = \lim_{\theta \to +\infty} \gamma_I(\theta) = \frac{1}{1 + D}.
\]
The first equality follows from the fact that \( \gamma_I \) is strictly increasing; the second from Lemma 1. Finally,
\[
\sup_{\theta \in \mathbb{R}} \gamma_I(\theta) = \lim_{\theta \to -\infty} \gamma_I(\theta) = \frac{1}{1 + D}.
\]
The latter equality comes by first applying Lemma 1 to all of the \( H_{I-1,s} \) and simplifying. This proves the first part of the theorem.

Now define \( \beta_{2,I} = (\beta_2, \ldots, \beta_I) \). Then,
\[
\gamma_I(\theta) = \frac{1}{1 + D} + \sum_{s=I}^{I+D} \left( \frac{1}{I - s + D} - \frac{1}{I - s + 1 + D} \right) \cdot H_{I-1,s}(\beta_{2,I}).
\]
Applying Lemma 2 to this expression gives the desired result. □

Proof of Corollary 1. The continuity of \( \gamma_I \) follows from the continuity of \( F_I \) and \( \gamma_I \). Choose \( \theta, \theta' \) with \( \theta > \theta' \). Then,
\[
G(\theta) - G(\theta') > \gamma_I(\theta) - \gamma_I(\theta') + [1 - \gamma_I(\theta)] \cdot [F_I(\theta) - F_I(\theta')]
\]
\[
> 0,
\]
since \( \gamma_I \) and \( F_I \) are strictly increasing. The range follows from combining the fact that \( G(\theta) \) is strictly increasing with \( \lim_{\theta \to -\infty} G(\theta) = 0 \) and \( \lim_{\theta \to +\infty} G(\theta) = 1 \). □

Proof of Theorem 2. We begin by computing \( \Pr(T_p = 0 \mid S_p = 0) \). For any \( A \subseteq \{1, \ldots, I\} \), define \( C_A \) to be the set of response patterns for which every item in \( A \) is answered correctly. Defining \( A \) to be the power set of \( \{1, \ldots, I\} \),
\[
\Pr(T_p = 0 \mid S_p = 0) = 1 - \Pr(T_p \geq 0 \mid S_p = 0)
\]
\[
= 1 - \sum_{A \subseteq A} (-1)|A| \Pr(C_A \mid S_p = 0)
\]
\[
= 1 - \sum_{A \subseteq A} (-1)|A| \cdot \frac{(I + D - |A|)!}{(I + D)!}
\]
\[
= 1 - \sum_{w=1}^{D} (-1)^{w+1} \cdot \frac{(I + D - w)!}{(I + D)!}
\]
where the second line follows from the Principle of Inclusion–Exclusion. By this argument, we can see that the probability of incorrectly answering every item of a set with \( m \) items and \( d \) distractors is
\[
\pi_m^d = 1 - \sum_{w=1}^{m} (-1)^{w+1} \cdot \frac{(m + d - w)!}{(m + d)!}.
\]
We now compute \( \Pr(T_p \mid S_p = 0) \). For any value of \( T_p \), there are \(\binom{I}{T_p} \) ways of selecting \( T_p \) correct items, the probability of each is \( (I - T_p + D)!/(I + D)! \). The probability of answering the remaining \( I - T_p \) items incorrectly is \( \pi_{I-T_p}^T \). Combining these probabilities yields
\[
\Pr(T_p \mid S_p = 0) = \left( \frac{I}{T_p} \right) \cdot \frac{(I - T_p + D)!}{(I + D)!} \cdot \pi_{T_p}^I.
\]
We can generalize this expression by noting that, since known items are answered correctly, the probability of correctly answering \( T_p \) items given that \( s \leq T_p \) items are known is equivalent to the probability of correctly answering \( T_p - s \) items correctly from a set of \( I - s \) items when no items are known. This means
\[
\Pr(T_p = t \mid S_p = s) = \frac{(I - s)!}{(I - t - D)!} \cdot \pi_{t-D}^I.
\]
and
\[
\Pr(T_p = t \mid \theta_p) = E_{\theta_p} \Pr(T_p = t \mid S_p = s)
\]
\[
= \sum_{s=0}^{I} \Pr(T_p = t \mid S_p = s) \Pr(S_p = s \mid \theta_p). \quad \Box
\]

Proof of Corollary 2. It is sufficient to show \( \Pr(T_p \leq t \mid \theta_p) \leq \Pr(S_p \leq t \mid \theta_p) \). By Theorem 2,
\[
\Pr(T_p \leq t \mid \theta_p) = \sum_{l=0}^{t} \sum_{s=0}^{I} \Pr(T_p = t \mid S_p = s) \Pr(S_p = s \mid \theta_p)
\]
\[
\sum_{s=0}^{r} \left[ \sum_{t=1}^{r} \Pr(T_p = t \mid S_p = s) \right] \cdot \Pr(S_p = s \mid \theta_p) \\
= \sum_{s=0}^{r} \Pr(T_p \leq r \mid S_p = s) \cdot \Pr(S_p = s \mid \theta_p) \\
\leq \Pr(S_p \leq r \mid \theta_p),
\]

since \(\Pr(T_p \leq r \mid S_p = s) \leq 1\), proving stochastic dominance. □

References