A Sequential IRT Model for Multiple-Choice Items and a Multidimensional Extension

Applied Psychological Measurement

2016, Vol. 40(4) 243–257

© The Author(s) 2016

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0146621616631518

apm.sagepub.com

Sien Deng¹ and Daniel M. Bolt¹

Abstract

For certain multiple-choice tests, it might be theorized that respondents evaluate response options in a stepwise fashion. Statistical models that assume such a process may compete against models that imply a process in which all response options are simultaneously compared, such as Bock's nominal response model. In this article, a sequential response model for multiple-choice items (SRM-MC) is considered. The model is applied to a sentence correction test in which the recognition of error and correction of error can be viewed as separate steps in solving an item. The proposed model permits the introduction of different proficiencies across steps. A fully Bayesian approach to estimating the model is presented, and an empirical comparison is performed against competing models. Empirical results support the proposed model and suggest distinct proficiencies related to recognition and correction.

Keywords

multiple-choice items, sequential response, sentence correction test

Many item response models have now been developed for multiple-choice items. A potentially useful distinction can be made between models that assume a divide-by-total structure (see Thissen & Steinberg, 1986) from those that are sequential (e.g., De Boeck & Partchev, 2012; Tutz, 1990). Divide-by-total models, including the nominal response model (NRM; Bock, 1972) and extensions (e.g., Revuelta, 2014; Samejima, 1979; Thissen, Cai, & Bock, 2010; Thissen & Steinberg, 1984), emphasize a single comparative evaluation of all response options that is influenced by one or more underlying latent traits. Such models emphasize a relationship between underlying latent trait(s) and an unobserved propensity toward each response option, with the chosen option influenced by the relative size of the propensities. Sequential models, by contrast, emphasize a stepwise response process, generally one in which particular response options are eliminated across steps, and can be represented using tree structures (see, e.g., Böckenholt, 2012b; De Boeck & Partchev, 2012; Jeon & De Boeck, 2015). To the extent that such tree structures are specified a priori, they impose restrictions on the assumed response process that can impact the statistical fit of model to test data, but to the benefit of potentially learning more about the process by which respondents are selecting among response options.

¹University of Wisconsin–Madison, USA

Corresponding Author:

Sien Deng, Department of Educational Psychology, University of Wisconsin–Madison, 1025 W. Johnson Street, Madison, WI 53706-1796, USA.

Email: sdeng7@wisc.edu

These two general modeling approaches can also be integrated. Examples included models where a divide-by-total model is used within a single step of a sequential model (see, e.g., Suh & Bolt, 2010), or where the same response can be achieved at different steps of the sequential process, such as in an extended cognitive miser model (Böckenholt, 2012a), or in the binary case, San Martín, del Pino, and De Boeck's (2006) ability-based guessing model. In Böckenholt's (2012b) cognitive miser model, a distinction is made between intuitive (immediate nondeliberative) responses and deliberative (thoughtful reasoning) responses, whereby an initial stage (described as "inhibitory control") must be passed before a deliberative response can be achieved. Extended versions of the cognitive miser model allow the same response to be attained at different stages of the response process.

The model considered in this article bears a close structural resemblance to an extended cognitive miser model. Some fundamental differences include (a) the modeling of multiple-choice response options in this article as opposed to open-ended responses in Böckenholt (2012a), (b) the use of slope parameters in the initial step of the model, and (c) a restriction of the first step toward consideration of a single response option. Importantly, the first difference makes it possible to compare the current model against well-defined competitor models that in turn permit stronger conclusions regarding a hypothesized response process. In the later presentation of the currently proposed model, other parallels with an extended cognitive miser model will be considered.

A useful feature of a sequential modeling structure is the potential to attach unique proficiency dimensions to separate steps in the sequential process. In certain contexts, this allows the model to capture misconceptions that may be associated with the selection of particular distracters, and statistically provides a mechanism by which dependencies in distracter selection across items can be explained. Attending to such misconceptions within a model also allows for quantification of how the misconception may plague performance across the test. It is possible that a test score loses its validity when a single misconception disproportionately penalizes an examinee across multiple items. The proposed model in relation to a sentence correction test is illustrated.

Real Data Illustration: A Sentence Correction Test

The real data application in this article relates to a sentence correction test that is administered as part of an English placement examination to entering students in a university system. A multiple-choice format is used for all items, and the items are scored as binary (1 = correct; 0 = incorrect). Each item has five response categories. The item stem always consists of an underlined portion of a sentence that potentially contains a grammatical error. Examinees are asked to select among five response options the best substitute for the underlined portion of the sentence without changing the meaning of the original sentence. A consistent item feature is that the first option of each item, option A, is always an exact replica of the underlined portion of the sentence, implying no error. Two example items are as follows:

Example Item 1. Heavy smoking and to overeat are activities which a heart patient must forego.

- A. Heavy smoking and to overeat
- B. Smoking heavily and to overeat
- C. To smoke heavily and overeating
- *D. Heavy smoking and overeating
- E. Smoking heavy and to overeat

Table 1. Descriptive Statistics of the Sentence Correction Test Total Scores.

Test	Sample	No. of items	Minimum	Maximum	M	SD
Sentence correction	5,848	23	1	23	12.27	3.69

Example Item 2. In the smaller towns of Michigan, where one can quickly walk to the greening hills of spring.

- A., where one can quickly walk
- B. where one can quickly walk
- C., where one can quickly walk,
- D. one can, quickly walk
- *E., one can quickly walk

For the above items, options “D and “E”, respectively, are the correct answers, as these represent the best substitutes. A variety of forms of actual errors are introduced on the test. These include errors related to the use of verb (e.g., subject–verb agreement, tense consistency), pronoun (e.g., consistency, case, agreement and reference), diction (e.g., idiom), modifier (e.g., adjective and adverb form and placement), and punctuation (e.g., fragment, comma fault, and punctuation for clarity). Similar sentence correction tests are used with many current standardized assessments, including the Scholastic Achievement Test (SAT) and the Graduate Management Admissions Test (GMAT), ACCUPLACER, the Praxis assessments, and several state tests, among others.

Data from 5,848 examinees to a 23-item form of the English sentence test were used in the current analysis. Table 1 displays descriptive statistics for correct total scores on the test. It is important to note that for the test form under consideration, the option “A” was the correct response for only 3 of the 23 items, specifically items 5, 14, and 19.

A Sequential Response Model for Multiple-Choice Items (SRM-MC)

Under the proposed model, a two-step sequential response process is specified by which an examinee arrives at a final chosen response option for each sentence correction item. The first step entails recognizing an error (or the possibility of an error) in the sentence. If no error is believed to exist in the sentence, then option A, a replica of the underlined portion of the sentence (implying “no error”), is selected. Alternatively, if the potential for an error is believed to exist, the examinee proceeds to a second step of considering a proposed correction. In this step, all five response options are evaluated, and the best alternative is selected. These two steps are referred to as the *recognition* and *correction* steps, respectively. To the extent that option A represents no proposed change, the response process can also be viewed as one in which the first response option is evaluated prior to considering the remaining options. It is important to note that the labeling of the steps as recognition and correction steps refers to the examinee’s potential action at each step (based on a perception of error) rather than the reality of an error. Consequently “passing” the recognition step only means that the examinee proceeds to the correction step, regardless of whether the sentence in actuality contains an error. For the three items on this test where the sentence actually contains no error, not passing the recognition step leads to a correct response.

It is possible that different examinee proficiencies are involved for the recognition and correction steps. Indeed, the specification of the proposed model was driven in part by an observation that examinees tend to vary substantially in the frequency with which option A is chosen as

a distracter and that such tendencies appeared imperfectly aligned with overall performance on the test (i.e., some well-performing examinees select option A with disproportionately high frequency). There are various possible explanations for why respondents may disproportionately select option A across items, including, for example, (a) a prior belief that a large number of items would be presented as correct, (b) rushed responding (as the test is a timed test), or (c) a high tolerance for incorrect grammar, among others. It will be assumed that all of these explanations are reflected through a poor recognition proficiency.

As shown below, the resulting model implies that an examinee can arrive at option A in two ways. First, the respondent may fail the recognition step and not proceed to evaluate other options. Alternatively, the respondent may pass the recognition step but after evaluating all five options consider option A to be better than the alternatives. Thus, the overall probability of selecting option A is viewed as the sum of probabilities associated with two pathways to that response. By contrast, in selecting options B through E, the only pathway is that the respondent first passes the recognition step and in the correction step views the option as the most preferred option among the five options. As noted earlier, this general modeling structure can be viewed as a form of the extended cognitive miser model of Böckenholt (2012a), but where a different interpretation is attached to the distinct steps. In the current application, the recognition step parallels an “inhibitory control” step of cognitive miser, and the correction step parallels a “deliberative reasoning step.”

Statistical Representation of the SRM-MC

Suppose a multiple-choice test is composed of n items and each item is of the format previously described, implying a total of m response categories per item, and p examinees. Let i index item ($i = 1, 2, \dots, n$) and j index examinee ($j = 1, 2, \dots, p$). In the recognition step, it is assumed that W_{ij} represents the evaluation of item i by examinee j such that $W_{ij} = 1$ implies that the sentence is recognized as incorrect (or possibly incorrect), a passing of the first step, while $W_{ij} = 0$ implies the sentence is perceived as correct, and option A is chosen. Furthermore, let U_{ij} denote the response to item i by examinee j , where $U_{ij} = k$ implies that category k is selected by examinee j for item i . In a two-dimensional application of the SRM-MC model, two different traits across steps are assumed. Specifically, η_j is denoted as the recognition tendency, while θ_j is the correction proficiency. Each step is modeled as follows.

Step 1: The probability of passing the recognition step is expressed as a traditional two-parameter logistic item response theory (2PL-IRT) model:

$$P(W_{ij} = 1 | \eta_j) = \frac{\exp(a_i \eta_j + d_i)}{1 + \exp(a_i \eta_j + d_i)}, \quad (1)$$

where a_i and d_i denote the slope and intercept parameters for item i .

Step 2: Assuming the recognition step is passed ($W_{ij} = 1$), the probability of selecting category k is expressed using Bock's NRM. Thus, the conditional probability of $U_{ij} = k$ given $W_{ij} = 1$ is modeled as

$$P(U_{ij} = k | W_{ij} = 1, \theta_j) = \frac{\exp[Z_{ik}(\theta_j)]}{\sum_{v=1}^{m_i} \exp[Z_{iv}(\theta_j)]}, \quad (2)$$

where a multivariate logit $Z_{ik}(\theta_j) = \lambda_{ik}\theta_j + \zeta_{ik}$ is used to define a propensity toward each category k as under the NRM, and arbitrary linear restrictions are imposed on the category parameters within items $\sum_{k=1}^{m_i} \lambda_{ik} = 0$ and $\sum_{k=1}^{m_i} \zeta_{ik} = 0$.

The overall probability that examinee j selects category k on item i can be achieved by combining results across the two steps. For the current application, it is assumed that $k = 1$ denotes “A,” $k = 2$ denotes “B,” up to $k = 5$, which denotes “E.” The overall probability model can be written as

$$P(U_{ij}=k|\theta_j, \eta_j) = I(U_{ij} = 1) \cdot P(W_{ij} = 0|\eta_j) + P(W_{ij} = 1|\eta_j) \cdot P(U_{ij} = k|W_{ij} = 1, \theta_j). \quad (3)$$

A unidimensional version of SRM-MC can be applied by replacing η_j with θ_j in equation (3), implying both steps are influenced by the same underlying trait. A fundamental question in applying the model is whether $\theta_j = \eta_j$ or the traits are distinct.

The item parameters of the recognition step of the SRM-MC (Equation 1) relate to aspects of the item that influence passing the recognition step. Items with larger more positive d parameters are items for which this step is easier to pass. The a parameter functions as a discrimination parameter implying that certain items may be more influenced by the recognition tendency than others.

Comparison Models for Multiple-Choice Items

To evaluate the plausibility of the proposed model for the sentence correction test, comparison models that provide alternative statistical representations that are aligned with alternative forms of response process are considered. Similar to the SRM-MC, each model can be written in unidimensional and two-dimensional forms. In the two-dimensional case, θ_j is referred to as the primary trait and η_j as a secondary trait.

NRM

A competing approach to the SRM-MC can be attained using the structure of the NRM. In unidimensional applications, the NRM defines a multinomial logit in reference to a single underlying latent trait. However, multidimensional extensions of the NRM have been considered, including Bolt and Johnson (2009), Thissen et al. (2010), and Revuelta (2014). By imposing appropriate constraints on a multidimensional NRM, the model can be used to define traits that represent over- or under-selection of particular response options, such as option A in the sentence correction test. Let $U_{ijk} = 1$ represent the selection of category k on item i by examinee j , and define a propensity $Z_{ik}(\theta_j, \eta_j) = \lambda_{ik1}\theta_j + \lambda_{ik2}\eta_j + \zeta_{ik}$ to be the attractiveness of category k on item i for examinee j , where θ_j and η_j represent different latent traits. The probability of selecting category k under the 2D-NRM is then expressed as

$$P(U_{ijk} = 1 | \theta_j, \eta_j) = \frac{\exp[Z_{ik}(\theta_j, \eta_j)]}{\sum_{v=1}^{m_i} \exp[Z_{iv}(\theta_j, \eta_j)]}, \quad (4)$$

where $\sum_{k=1}^{m_i} Z_{ik}(\theta_j, \eta_j) = 0$, indicating $\sum_{k=1}^{m_i} \lambda_{ik1} = 0$, $\sum_{k=1}^{m_i} \lambda_{ik2} = 0$, and $\sum_{k=1}^{m_i} \zeta_{ik} = 0$.

To make the NRM account for a similar trait (η_j) to the recognition tendency under the SRM-MC, constraints can be applied to the λ_{ik2} on the η_j dimension. Specifically, by specifying as category slopes, $\lambda_{i2} = (\lambda_{i2}, -.25\lambda_{i2}, -.25\lambda_{i2}, -.25\lambda_{i2}, -.25\lambda_{i2})$, where λ_{i2} is a single slope parameter to be estimated with respect to the second dimension for each item, η_j represents a respondent's tendency toward selecting option A. Allowing λ_{i2} to vary across items permits this tendency to more heavily influence responses to certain items compared with others, similar to how the a_i in the SRM-MC model functioned as a discrimination parameter for η_j in regard to the recognition step.

A fundamental difference between the NRM and SRM-MC relates to their assumptions regarding response process. For the NRM, the trait η_j is simply an added factor present in the comparative evaluation of all response options. Under the SRM-MC model, η_j influences whether one proceeds to a comparison step. As a result, the SRM-MC provides an explanation for why a respondent may select option A even in the presence of highly attractive alternative distracters. The stepwise approach underlying the SRM-MC assumes that in many instances the alternative response options are simply not evaluated by the examinee. Thus, seeing a statistical difference in fit between the two models can lend some insight into the underlying response process.

A unidimensional version of the NRM can also be applied for comparative purposes and is specified as

$$P(U_{ijk} = 1 | \theta_j) = \frac{\exp[Z_{ik}(\theta_j)]}{\sum_{v=1}^{m_i} \exp[Z_{iv}(\theta_j)]}, \quad (5)$$

where $Z_{ik}(\theta_j) = \lambda_{ik}\theta_j + \zeta_{ik}$, $\sum_{k=1}^{m_i} \lambda_{ik} = 0$, and $\sum_{k=1}^{m_i} \zeta_{ik} = 0$.

The unidimensional NRM naturally assumes that only one underlying trait influences the selection of response options across items.

Nested Logit Model (NLM)

A second modeling approach considered for comparison purposes is a two-dimensional NLM (2D-NLM, Bolt, Wollack, & Suh, 2012). The NLM has a similar structure to the SRM-MC but instead models the probability of distracter selection conditional upon an incorrect response. Let $X_{ij} = 1$ represent a response to item i by examinee j that is correct, and $D_{ijk} = 1$ denote that examinee j selects distracter category k on item i . Under the NLM, the probability that an examinee with ability θ_j selects the correct category on item i is modeled by the 2PL-IRT model as

$$P(X_{ij} = 1 | \theta_j) = \frac{\exp(a_i\theta_j + b_i)}{1 + \exp(a_i\theta_j + b_i)}, \quad (6)$$

where b_i is an intercept parameter and a_i a slope parameter for item i .

The probability that an examinee selects distracter category v on item i is modeled as the product of the probability of an incorrect response and the probability of choosing distracter v conditional on an incorrect response under Bock's NRM. In the 2D-NLM, a different proficiency can be similarly introduced to the conditional probability relative to the proficiency assumed to underlie the correctness of the response. The probability of selecting a particular response option is then modeled as

$$\begin{aligned}
 P(X_{ij} = 0, D_{ijk} = 1 | \theta_j, \eta_j) &= P(X_{ij} = 0 | \theta_j) P(D_{ijk} = 1 | X_{ij} = 0, \eta_j) \\
 &= \left[1 - \frac{\exp(a_i \theta_j + b_i)}{1 + \exp(a_i \theta_j + b_i)} \right] \cdot \frac{\exp[Z_{ik}(\eta_j)]}{\sum_{v=1}^{m_i-1} \exp[Z_{iv}(\eta_j)]},
 \end{aligned} \tag{7}$$

where η_j denotes the trait that influences selection among the $m_i - 1$ distracters, and $Z_{iv}(\eta_j) = \zeta_{iv} + \lambda_{iv}\eta_j$.

For the sentence correction test, the NLM may represent a condition where prior to examining any of the response options, the examinee attempts to determine how the underlined portion of the sentence should be changed and looks for that proposed change among the list of the provided response options. If that specific correction is not present, the examinee would then engage in a comparative evaluation of response options. It is theorized that such a response process was less plausible for the sentence correction test. In particular, it would not likely provide an account for why option A is chosen more frequently than the other response options by certain examinees. As with the NRM, a unidimensional NLM can be viewed as a special case of the two-dimensional NLM by setting $\theta_j = \eta_j$.

Fully Bayesian Estimation Algorithms

For both the SRM-MC and comparison models, a fully Bayesian approach was implemented for the estimation of model parameters. Markov Chain Monte Carlo (MCMC) methods were applied using WinBUGS 1.4 (Lunn, Thomas, Best, & Spiegelhalter, 2000). Based on the specified model and associated prior distributions for item parameters, parameter states are sampled in a fashion that permits estimation of the multivariate posterior distribution of model parameters. For each of the models under consideration, a Metropolis-Hastings sampling algorithm was used. Five simulated Markov chains were run out to 10,000 iterations. An initial run of 4,000 iterations was used to define proposal distributions, and parameter estimates were calculated based on the subsequent 6,000 iterations.

For the SRM-MC, the prior distributions of model parameters were set at levels commonly applied with related IRT models (Bolt et al., 2012; Patz & Junker, 1999).

Specifically, for the item parameters it is assumed that

$$a_i \sim \log \text{Normal}(0, 1),$$

$$b_i \sim \text{Normal}(0, 1),$$

$$\lambda_{ik} \sim \text{Normal}(0, 1), \text{ and}$$

$$\zeta_{ik} \sim \text{Normal}(0, 1),$$

while for the person parameters, it is assumed that

$$\theta, \eta \sim \text{BivNormal}(0, \Sigma),$$

where

$$\Sigma = \begin{pmatrix} 1 & \rho_{\theta, \eta} \\ \rho_{\theta, \eta} & 1 \end{pmatrix}, \quad \rho_{\theta, \eta} \sim \text{Uniform}(-1, 1).$$

To make the NRM and NLM comparable to the SRM-MC, the priors used for these comparison models largely match those above. Priors for person parameters in the two-dimensional NRM and NLM are the same as those used for the two-dimensional SRM. Priors for item parameters are also similar to those in the two-dimensional SRM. For the two-dimensional NRM,

$$\lambda_{ik1} \sim \text{Normal}(0, 1),$$

$$\lambda_{i2} \sim \text{Normal}(0, 1), \text{ and}$$

$$\zeta_{ik} \sim \text{Normal}(0, 1),$$

and for the two-dimensional NLM,

$$a_i \sim \log \text{Normal}(0, 1),$$

$$b_i \sim \text{Normal}(0, 1),$$

$$\lambda_{ik} \sim \text{Normal}(0, 1), \text{ and}$$

$$\zeta_{ik} \sim \text{Normal}(0, 1).$$

Unidimensional versions of the models used the same item parameter priors as the two-dimensional models, and for the person parameters set $\theta \sim \text{Normal}(0, 1)$.

The SRM-MC, NRM, and NLM models contain varying overall numbers of parameters. To statistically compare models, a penalty related to the number of parameters is taken into account. In a Bayesian estimation framework, the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & Van Der Linde, 2002) is an information-based model comparison index that is commonly used. The DIC is calculated as $\text{DIC} = p_D + \bar{D}$, where p_D is the effective number of parameters in the model, and $\bar{D}$ is the mean deviance, reflecting the average deviance when integrating over the parameter space. A model with better comparative fit returns a lower DIC.

Results

All models demonstrated good convergence properties with respect to the simulated Markov chains. Although the use of MCMC algorithms with unidimensional and two-dimensional NRM and NLM models has been studied in other articles (e.g., Bolt et al., 2012), the use of these methods with the SRM-MC model is new. Among other criteria, the potential scale reduction factor (PSRF) of Gelman and Rubin (1992) is examined using the five simulated chains out to 10,000 iterations. Virtually all of the parameters were below the recommended level of 1.1. Specifically, values below 1.1 were observed for 23/23 of the a parameters, 22/23 of the d parameters, 115/115 of the λ parameters, and 112/115 of the ζ parameters. Comparable convergence results were observed with respect to the examinee parameters. Despite good convergence results, the average posterior standard deviation (psd) for $\hat{\eta}$ was 0.811, suggesting some difficulty in estimating this parameter at the individual examinee level. Higher psds tended to be observed for examinees with higher $\hat{\theta}$ due to the smaller numbers of incorrect responses.

Model Comparison Results

Table 2 displays model comparison results for unidimensional and two-dimensional versions of each of the models. For the NRM and SRM-MC models, the two-dimensional models appear to

Table 2. DIC Model Comparison Results for Sentence Correction Test.

Model	$\bar{D}$	$\hat{D}$	p_D	DIC
Unidimensional SRM-MC	285,675	281,286	4,389	290,065
Unidimensional NRM	285,625	281,153	4,472	290,097
Unidimensional NLM	285,398	280,884	4,514	289,911
Two-dimensional SRM-MC	282,502	276,075	6,427	288,929
Two-dimensional NRM	282,419	275,842	6,577	288,996
Two-dimensional NLM	284,499	279,064	5,435	289,933

Note. DIC = deviance information criterion; SRM-MC = sequential response model for multiple-choice items; NRM = nominal response model; NLM = nested logit model.

Table 3. Correlations Between Ability Estimates $\hat{\theta}$ From Comparable Models.

$\hat{\theta}$ correlations	2D-NRM	2D-NLM
2D-SRM	.994	.990
2D-NRM		.996

Note. NRM = nominal response model; NLM = nested logit model; SRM = sequential response model.

fit better than their unidimensional counterparts. For the NLM model, the unidimensional version appears slightly preferred to the two-dimensional model. Interestingly, among unidimensional models, the NLM appears to be the preferred model. The observation that the unidimensional model would be preferred to the two-dimensional model under the NLM importantly highlights how dimensionality can also be influenced by the specific structure assumed by the item response model.

Across all six models, the two-dimensional SRM-MC provides the best comparative fit. As noted, the two-dimensional NRM was viewed as providing the most meaningful comparison in terms of response process. The DIC estimates for the two-dimensional SRM-MC and NRM are quite close, with a slight preference for the SRM-MC, suggesting the response process represented by the SRM-MC may be more consistent with the sentence correct test. Such findings support the notion that certain examinees are more prone to select option A by not evaluating alternative response options, as this is the primary distinguishing structural feature of the SRM-MC and NRM models.

A further comparison of models can occur by correlating their resulting proficiency estimates. Tables 3 and 4 show the correlations between the examinee proficiency estimates across the two-dimensional versions of each model. In each case, the proficiency estimate is obtained as the mean of the sampling history for the corresponding person parameter (omitting burn-in iterations) in the MCMC estimation of each model. In Table 3, it is seen that the primary latent trait θ (representing the correction proficiency in the SRM, response correctness in the NRM and NLM) yields correlations that are all at or above .99 across all models, suggesting that the same primary trait appears to be measured more or less in a consistent fashion by all models. Results for the secondary trait η in Table 4 (representing recognition tendency in the SRM, overselection of option A in the NRM, and a trait underlying distracter selection in the NLM) suggest a close correspondence between these secondary traits for the SRM-MC and NRM, which are correlated above .90. The high correlation is expected as the constraints applied to the NRM were chosen to make this trait similar to that of SRM-MC. The somewhat weaker

Table 4. Correlations Between Ability Estimates $\hat{\eta}$ From Different Two-dimensional Models.

$\hat{\eta}$ correlations	2D-NRM	2D-NLM
2D-SRM	.919	.296
2D-NRM		-.020

Note. NRM = nominal response model; NLM = nested logit model; SRM = sequential response model.

Table 5. Correlations Between $\hat{\theta}$ and $\hat{\eta}$ From Different Two-Dimensional Models.

Models	$\hat{\rho}_{\theta,\eta}$
2D-SRM	-.233
2D-NRM	.041
2D-NLM	.824

Note. SRM = sequential response model; NRM = nominal response model; NLM = nested logit model.

correlation between these traits across models in comparison with the primary proficiencies could be due to the reduced reliability of the secondary trait estimates. At the same time, there is little to no relationship between secondary traits for the NLM in relation to the other models. This is also as expected, given that the NLM is attending to a latent trait influencing selection among distracters conditional upon an incorrect response. It might be anticipated that this trait would more closely align with the primary proficiency than occurs under the competing models, which attend to option A specifically. Table 5 suggests this to be the case, as θ and η in the NLM are seen to be highly positively correlated. However, for the SRM-MC and NRM, the low correlation between θ and η implies a secondary trait quite different from the primary trait.

Results of the Sequential Response Model (SRM-MC)

Given its superior performance in the model comparison study, the results observed for the SRM-MC are considered in greater detail. As the item response process underlying the sentence correction items appears to more closely follow the response process assumed by the new model, the person and item parameter estimates may provide further insights into both individual examinees and the functioning of test items.

As noted above, a unique feature of the SRM-MC is its capacity to estimate an examinee's disproportionate tendency toward selecting option A. This feature can be seen from the example response patterns presented in Table 6, which shows response patterns and corresponding $\hat{\theta}$ and $\hat{\eta}$ for nine examinees. The first three examinees represent examples displaying low recognition proficiency estimates ($\hat{\eta}$), while the last three examinees have much higher estimates. In each case, the recognition tendency estimate appears closely related to the overall frequency with which option A is selected, with lower $\hat{\eta}$ s corresponding to more frequent selection of option A.

Importantly, one of the characteristics that make the first examinee a low $\hat{\eta}$ examinee is a tendency to select option A even in the presence of highly attractive alternative distracter options. For example, this examinee selects option A as an incorrect response on items 4, 13, and 23, each of which is an item for which the overall frequency of selecting category A is very low (see Table 8), and much more attractive distracters exist. Under the NRM, such responses will be very low-likelihood responses; however, in the SRM, such responses can be explained

Table 6. Example Response Patterns and SRM-MC Proficiency Estimates.

ID	Response	Total score	No. of category A	$\hat{\theta}$	$\text{psd}(\hat{\theta})$	$\hat{\eta}$	$\text{psd}(\hat{\eta})$
1	CAAAABABAABDAAAAACAAAAA	4	17	-2.325	0.779	-3.413	0.716
2	BAADABABAABAEABABDAAE	8	12	-0.571	0.659	-2.715	0.721
3	BDEAADACAADDEAEBAABDAD	16	9	1.638	0.692	-1.299	0.718
4	BACACBCEBBDAEAEBCBCBB	5	5	-2.503	0.510	-0.61	0.751
5	BDCDCBBBACADBCDEBDABDAD	10	4	0.001	0.497	0.034	0.787
6	BDADADDCBADDEAEEDDADDAD	19	6	1.783	0.640	-0.583	0.710
7	EDCBECECCBDBCBBBEBEDBCD	6	0	-1.626	0.448	0.765	0.858
8	DDCCCBDBBDBBCEDEBDBCECD	10	0	-0.37	0.481	0.916	0.768
9	BDEDADDCBDDCEAEEDBADDCCD	22	0	2.562	0.703	1.235	0.908

Note. SRM-MC = sequential response model for multiple-choice items; psd = posterior standard deviation.

by a low $\hat{\eta}$, suggesting the examinee simply failed to see a possible error in the sentence (and thus did not evaluate alternatives).

As noted earlier, the emergence of $\hat{\eta}$ as a dimension in the data distinct from $\hat{\theta}$ has several potential interpretations. It may well reflect a unique aspect of English usage proficiency, namely, the ability to recognize a problematic sentence. But it could also reflect more of a nuisance dimension to the extent that the rules of English usage are not always well defined and subject to different interpretations. Regardless of which is correct, there will likely be value in understanding the relative contributions of these two traits to overall performance on the test. There are different ways in which such effects can be evaluated. On one hand, a very high correlation ($>.97$) is observed between the $\hat{\theta}$ estimated under a standard unidimensional NRM and the $\hat{\theta}$ seen under the two-dimensional sequential response model, suggesting that on average, the primary proficiency estimates are not affected much. On the other hand, such a high correlation does not preclude seeing effects of $\hat{\eta}$ in relation to individual examinee scores. To examine this issue, a regression model is applied attempting to predict test score as a function of $\hat{\theta}$ and $\hat{\eta}$ estimates from the SRM-MC model. From the real data analysis, $\text{sd}(\hat{\theta}) = 0.876$ and $\text{sd}(\hat{\eta}) = 0.634$. Results from the regression analysis are shown in Table 7.

From this analysis, it can be seen that both $\hat{\theta}$ and $\hat{\eta}$ are relevant to test performance. Importantly, each unit change of $\hat{\eta}$ in the negative direction implies essentially one additional score (i.e., 1.009) point lost. For the first three examinees in Table 6, for example, the consequence of their extreme $\hat{\eta}$ scores would be a loss of approximately 3 score points on the test. If $\hat{\eta}$ were viewed as a nuisance dimension, such results could be the basis for proposed score changes that would neutralize the effect of $\hat{\eta}$ on test performance. Even if such a correction were not applied, the model estimates give a clearer sense as to the consequence of this proficiency on test scores, in this case with θ being about 3.5 times as consequential as η . Interestingly, a corresponding analysis using the NRM estimates suggests an influence of η that is only about 2/3 the magnitude of that implied by the SRM-MC, suggesting one practical difference the SRM-MC and NRM approaches concerns the relative influence of $\hat{\eta}$ on test performance.

Application of the SRM-MC model also permits study of the item characteristics in relation to the recognition and correction steps of the model. The focus is in particular on differences across items in relation to η and the recognition step. Table 8 shows the item slope (a -parameter), intercept (d -parameter) and item difficulty (b -parameter) under the SRM-MC, the corresponding overall proportions of examinees that selected option A on each item, as well as the

Table 7. Regression Results Predicting Test Score as a Function of $\hat{\theta}$ and $\hat{\eta}$ Using SRM-MC Estimates.

Predictor	<i>b</i>	SE (<i>b</i>)	<i>t</i>	<i>p</i> value
Intercept	12.221	.017	713.399	<.001
$\hat{\theta}$	3.597	.021	170.163	<.001
$\hat{\eta}$	1.009	.029	34.562	<.001

Note. SRM-MC = sequential response model for multiple-choice items.

Table 8. Item Discrimination and Difficulty Involved in Step 1 Only, the Proportion Selecting Option A, and Proportion Correct.

Item	$\hat{a}$ (psd)	$\hat{d}$ (psd)	$\hat{b} = -\frac{\hat{d}}{\hat{a}}$	% category A	% correct response (delta metric)
1	0.618 (.195)	4.598 (.300)	-7.439	2.2	0.58
2	0.563 (.192)	2.282 (.471)	-4.056	20.6	0.55
3	0.812 (.133)	2.210 (.265)	-2.723	16.4	-0.59
4	0.665 (.152)	3.972 (.207)	-5.977	3.7	1.01
5	0.384 (.114)	2.238 (.182)	-5.822	33.1	-0.44
6	0.211 (.108)	2.166 (.445)	-10.285	17.1	-0.28
7	0.455 (.114)	2.148 (.273)	-4.719	14.8	-0.03
8	0.447 (.178)	4.691 (.256)	-10.492	1.7	0.30
9	0.671 (.204)	1.992 (.474)	-2.967	25.0	0.20
10	1.352 (.160)	1.785 (.215)	-1.320	31.8	0.07
11	0.336 (.116)	2.313 (.290)	-6.892	20.0	-0.13
12	0.691 (.160)	2.821 (.343)	-4.081	11.6	0.69
13	0.342 (.158)	4.308 (.272)	-12.611	2.6	0.55
14	0.788 (.137)	1.675 (.233)	-2.127	42.9	-0.18
15	1.025 (.109)	2.242 (.159)	-2.187	16.8	-0.27
16	0.483 (.178)	2.645 (.404)	-5.477	15.5	0.64
17	0.226 (.085)	2.045 (.238)	-9.053	15.9	0.38
18	0.862 (.150)	2.817 (.292)	-3.270	16.2	0.21
19	0.605 (.149)	1.679 (.270)	-2.774	57.4	0.18
20	0.455 (.130)	2.421 (.304)	-5.322	13.7	-0.73
21	0.974 (.219)	3.406 (.437)	-3.498	12.0	-0.09
22	1.001 (.109)	0.929 (.160)	-0.928	37.2	-0.28
23	0.413 (.197)	5.419 (.295)	-13.127	1.0	-0.20

Note. psd = posterior standard deviation.

overall item difficulty (on the delta metric). Not surprisingly, the likelihood of passing the recognition step relates to the probability of an overall correct response as well as the probability of selecting option A. Moreover, it seems clear from the varying discrimination estimates that the items are differentially sensitive to $\hat{\eta}$, implying some items tend to evoke a sensitivity to accept the sentence as correct more than others. Interestingly, such effects appear unrelated to the item location in the test, suggesting rushed responding is likely not the explanation for this effect. Items displaying the largest $\hat{a}$ estimates, namely, items 10, 15, and 22, are uniquely items for which the problem with the sentences is with the punctuation; moreover, all are sentences in which the proposed use of a comma to separate clauses of the sentence should have been replaced with either a semicolon or dash. Such a finding suggests that the recognition proficiency is heavily related to an examinee's sensitivity to punctuation errors. It also introduces a possibility that the multidimensionality detected in this assessment might relate as much to

Figure 1. Scatterplot of intercept estimate from the recognition step against the item proportion correct on the delta metric.

the types of errors (punctuation versus word usage) introduced as it does to a distinct recognition tendency.

As noted earlier, under the SRM model, an examinee can arrive at option A in two ways, by not passing the recognition step, or by passing the recognition step but finding option A to be the most plausible response in the correction step. Not surprisingly, the estimated intercepts related to the recognition step are rather highly correlated with the category intercept estimates of option A in the correction step ($r = .77$), implying that items for which option A is in general a plausible distracter (or the correct response) are also items for which passing the recognition step occurs with less frequency. Figure 1 shows the relationship between negative intercept in the “recognition step” and the overall proportion correct on the delta scale. The clear negative relationship reflects that items that are in general more difficult also tend to be more difficult with respect to the recognition step. The correlation between the intercept of the recognition step and the item difficulty is $-.364$, a relationship that is made even stronger ($-.520$) by removing item 23. It would thus seem that the ability to recognize a sentence as being problematic is one important component associated with an item’s difficulty level. Items that depart from this linear relationship often reflect items for which a particular alternative distracter was found highly attractive. For example, Item 23, which is a difficult item but with a low difficulty in regard to the recognition step, introduces an easily identified word usage problem, but one for which the appropriate correction is less clear.

Discussion and Conclusions

The sentence correction test analyzed in the article is frequently described as measuring a student's ability to simultaneously recognize and correct sentence errors. The model considered in this article provides a method by which to statistically evaluate whether the traits underlying recognition and correction are in fact the same. By finding a better comparative fit for the two-dimensional SRM-MC model in comparison with both the NRM and unidimensional SRM-MC models, it would appear that a theorized response process in which examinees are in many cases failing to recognize an error in the sentence and selecting option A (without evaluation of alternatives) is plausible; moreover, it would appear that the recognition and correction traits that emerged in this analysis are quite distinguishable. The recognition tendency that emerged from this analysis appears particularly related to identification of punctuation (as opposed to word usage) errors in sentences. As punctuation is often a somewhat subjective determination, it is perhaps not surprising that items with punctuation errors emerge as the most difficult and discriminating with respect to the recognition proficiency.

As emphasized in the introduction, the model considered in this article can be viewed as a form of the extended cognitive miser model (Böckenholt, 2012a). When applied to multiple-choice items, the model provides a statistical mechanism by which a theorized sequential process underlying the evaluation of response options can be tested against an alternative model that assumes simultaneous evaluation of all response options. A sequential process appeared reasonable for the current test based on the consistent way in which option A was specified across items and the logical sequence in which the first response option would be evaluated in relation to the subsequent response options. Variations on this model could also be developed in other contexts. An alternative assessment where a similar type of model structure could be applied might be a medical practitioner certification test, where items correspond to decision-making scenarios and where incorrect decisions can be made based on limited amounts of information provided to the prospective practitioner. Recent innovations in computerized assessment has made the design and use of such item types attractive and plausible.

As seen in the regression analysis, the SRM-MC model can also provide a way of evaluating how the recognition and correction traits ultimately impact examinee scores on the test. From this analysis, it would appear that the correction proficiency carries approximately 3½ times the weight of the recognition tendency. To the extent that parallel measures of the test are desired, evaluation of the relative weighting of these traits across test forms may be a useful criterion to consider.

It is conceivable that the trait η , although labeled as a recognition tendency, actually represents an unintended dimension of the test, in which case one consequence of these findings may be toward a suggested redesign of sentence correction test items. One possibility would be to mix the "same sentence" option throughout the response options across items, or perhaps consistently place it as the last response option, so as to encourage examinees to consider all response options. This may be a particularly important consideration if the relative influence of η is seen to vary significantly across test forms. Alternatively, examinees might be informed prior to the test about the approximate number of sentences on the test that contain no error.

Finally, there are a number of additional considerations regarding the use of the SRM-MC in the specific application considered in this article. Other approaches to validating the distinct recognition and correction steps in the SRM-MC model could be applied. For example, having think-aloud studies or post-test interviews with respondents could help in validating the presence of the two steps. In addition, attending to response times in the context of computerized assessments (the current analysis were based on paper-pencil test administrations) may

demonstrate shorter response times for many of the “A” responses, a result that would lend further support to the SRM-MC.

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) received no financial support for the research, authorship, and/or publication of this article.

References

- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29-51.
- Böckenholt, U. (2012a). The cognitive-miser response model: Testing for intuitive and deliberate reasoning. *Psychometrika*, 77, 388-399.
- Böckenholt, U. (2012b). Modeling multiple response processes in judgment and choice. *Psychological Methods*, 17, 665-678.
- Bolt, D. M., & Johnson, T. R. (2009). Applications of a MIRT model to self-report measures: Addressing score bias and DIF due to individual differences in response style. *Applied Psychological Measurement*, 33, 335-352.
- Bolt, D. M., Wollack, J. A., & Suh, Y. (2012). Application of a multidimensional nested logit model to multiple-choice test items. *Psychometrika*, 77, 339-357.
- De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, 48, 1-28.
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457-511.
- Jeon, M., & De Boeck, P. (2015). A generalized item response tree model for psychological assessments. *Behavior Research Methods*. Advance online publication. doi:10.3758/s13428-015-0631-y
- Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). Winbugs—A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10, 325-337.
- Patz, R. J., & Junker, B. W. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, 24, 342-366.
- Revue, J. (2014). Multidimensional item response model for nominal variables. *Applied Psychological Measurement*, 38, 549-562.
- Samejima, F. (1979). *A new family of models for the multiple choice item* (Research Report No. 79-4). Knoxville: Department of Psychology, University of Tennessee.
- San Martin, E., del Pino, G., & De Boeck, P. (2006). IRT models for ability-based guessing. *Applied Psychological Measurement*, 30, 183-203.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 64, 583-639.
- Suh, Y., & Bolt, D. M. (2010). Nested logit models for multiple-choice item response data. *Psychometrika*, 75, 454-473.
- Thissen, D., Cai, L., & Bock, R. D. (2010). The nominal categories item response model. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models: Developments and applications* (pp. 43-75). New York, NY: Taylor & Francis.
- Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika*, 49, 501-519.
- Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, 51, 567-577.
- Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, 43, 39-55.