PSYCHOMETRIKA
https://doi.org/10.1007/s11336-022-09880-8

ACCURATE ASSESSMENT VIA PROCESS DATA

SUSU ZHANG

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

ZHI WANG

CITADEL SECURITIES

JITONG QI, JINGCHEN LIU AND ZHILIANG YING

COLUMBIA UNIVERSITY

Accurate assessment of a student's ability is the key task of a test. Assessments based on final responses are the standard. As the infrastructure advances, substantially more information is observed. One of such instances is the process data that is collected by computer-based interactive items and contain a student's detailed interactive processes. In this paper, we show both theoretically and with simulated and empirical data that appropriately including such information in the assessment will substantially improve relevant assessment precision.

Key words: Process data, ability estimation, automated scoring, Rao–Blackwellization.

The main task of educational assessment is to provide reliable and valid estimates of test takers' ability based on their responses to test items. Much of the efforts in the past decades focused on the item response theory (IRT) models, the responses of which are often dichotomous (correct/incorrect), polytomous (partial score), or nominal (e.g., response choice). The rapid advancement of information technology has enabled the collection of various sorts of process data from assessments, ranging from reaction times on multiple-choice questions to the log of problem-solving behavior on computer-based constructed-response items. In particular, the sequence of actions performed by test takers to solve a task, which documents the problem-solving process, can contain valuable information on top of final responses, that is, dichotomous or polytomous scores on how well the task was completed. The analysis of process data has recently gained strong interest, with a wide range of model- and data-driven methods proposed to understand the types of strategies that contribute to successful/unsuccessful problem-solving, identify the behavioral differences between observed and latent subgroups, and assess the proficiency on the trait of interest, et cetera (e.g., He & von Davier, 2016; LaMar, 2018; Liu et al., 2018; Xu et al., 2018).

Process data often contain rich information about the test takers. Much of the literature focuses on developing new research directions. In this paper, we try to answer the question of how existing research could benefit from the analysis. Specifically, we develop a method to incorporate the information in process data into the scoring formula. There are two key features to consider in measurement, reliability and validity. The current paper focuses on the improvement in reliability: We show that the process-incorporated latent ability estimate improves measurement accuracy. In particular, we demonstrate through simulation and empirical analyses that the process-incorporated scoring rule yields much higher reliability than the IRT-model-based scoring rule. On

Correspondence should be made to Jingchen Liu, Columbia University, New York, NY, USA. Email: jcliu@stat.columbia.edu

Published online: 13 August 2022

the current empirical data. score based on two items' process data on average could be as accurate as that based on final scores on five items. Furthermore, we provide a theoretical framework under which process-incorporated scores yield lower measurement error of test takers' abilities under certain regularity conditions. One particular challenge in process-incorporated scoring is that process data record the entire problem-solving process, are highly unstructured and variable, and reveal different aspects of a test taker, including construct-irrelevant characteristics. It is unclear which part of the process is related to the particular trait of interest. Conventionally, it is up to the domain experts to identify construct-relevant process features and derive scoring rules. The proposed approach, on the other hand, considers an automated method: Features are extracted through an exploratory analysis that typically lacks interpretation. We take advantage of the IRT score, which serves as a guide for the scoring rule to yield an estimator for the measured trait. The entire procedure does not require particular knowledge of the item design.

Process data are often in a nonstandard format difficult to directly analyze. We preprocess the data by embedding each response process to a finite-dimensional vector space. There are multiple methods to fulfill this task, including n-gram language modeling (He & von Davier, 2016), sequence-to-sequence autoencoders (Tang et al., 2021b), and multidimensional scaling (Tang et al., 2021a). In this paper, we use features extracted from multidimensional scaling.

In the psychometrics literature, plenty of research has been conducted on the use of additional information, such as response times, to improve measurement accuracy (e.g., Bolsinova & Tijmstra, 2018). To the authors' best knowledge, this is the first piece of work to use problem-solving log data to improve measurement accuracy. A literature that is remotely related to the current work is automated scoring systems for constructed responses. Extensive research has been conducted on automated scoring of essays, which aims at producing essay scores comparable to human scores based on examinees' written text (e.g., Attali & Burstein, 2006; Foltz et al., 1999; Page, 1966; Rudner et al., 2006). Other than essay scoring, automated scoring engines have been developed for scenario-based questions in medical licensure examinations (Clauser et al., 2000), constructed-response mathematics problems (Fife, 2013), speaking proficiency examinations (Evanini et al., 2015), and prescreening for post-traumatic stress disorder based on self-narratives (He et al., 2017, 2019). Many of these systems were shown to produce comparable scores to expert ratings. Readers are referred to Bejar et al. (2016) and Rupp (Rupp (2018)) for comprehensive reviews of the history, applications, conceptual foundations, and validity considerations of automated scoring systems.

The proposed approach differs from most automated scoring systems in its objective. Whereas automated scoring systems are often designed to reproduce expert- or rubric-derived scores in an automated and standardized manner, the purpose of the proposed Rao–Blackwellization approach is not to reproduce the final scores but to refine the latent trait estimates based on original final scores with the additional information from the problem-solving process.

The rest of the paper is organized as follows. Section 1 describes the statistical formulation. The proposed method for process-incorporated score refinement is introduced. Theoretical results on mean squared error (MSE) reduction in latent trait estimation are presented. Section 2 reports the results of simulation studies that verify the theoretical findings. Section 3 presents an empirical example on the problem-solving in technology-rich environments (PSTRE) assessment in the 2012 Programme for the International Assessment of Adult Competencies (PIAAC) survey, where the proposed method is compared to original response-based scoring in several aspects. A discussion of the implications and limitations is provided in Sect. 4.

1. Latent Trait Estimation with Processes and Responses

This section presents a generic framework for process-incorporated latent trait measurement. Section 1.1 describes a statistical formulation that the current framework is built upon. Section 1.2 introduces the proposed Rao–Blackwellization approach and the specific procedures. Section 1.3 provides a theoretical analysis of the process-incorporated trait estimator.

1.1. Statistical Formulation

Consider a test of J items designed to measure a latent trait, θ. For an examinee, on each item j, both the final item response and the action sequence for problem-solving are recorded. Denote the item response by Y_j, which can be a polytomous score ranging between 0 and C_j, representing different degrees of task completion. Further denote the action sequence by $\boldsymbol{S}_j = (S_{j1}, \ldots, S_{jL_j})$, where L_j is the total number of actions performed on the item, and S_{jl} is the lth action.

We consider the case where the action sequences record problem-solving details and thus contain at least as much information as the final outcomes. In this case, the final item response can be derived from the action sequence through a deterministic scoring rule f such that $Y_j = f(\mathbf{S}_j)$. Further suppose that the final responses to the J items are conditionally independent given θ and follow some item response function (e.g., Lord, 1980)

$$P(Y_j = y_j \mid \theta, \boldsymbol{\zeta}_j),$$

where $\boldsymbol{\zeta}_j$ is the parameters of item j.

For the present purpose of latent trait estimation, we assume that the item parameters ($\boldsymbol{\zeta}_j$s) have been calibrated and only the latent trait θ is unknown. Denote the pre-calibrated parameters of item j by $\hat{\boldsymbol{\zeta}}_j$. The latent trait θ for each individual can be estimated based on the response from one or more items. Commonly used latent trait estimators include the maximum likelihood estimator (MLE),

$$\hat{\theta}^{\text{MLE}} = \underset{\theta}{\operatorname{argmax}} \sum_j \log(P(Y_j = y_j \mid \theta, \hat{\boldsymbol{\zeta}}_j)), \tag{1}$$

the Bayesian expected a posteriori estimator (EAP), and the Bayesian modal estimator (BME), i.e.,

$$\hat{\theta}^{\text{EAP}} = E[\theta \mid \mathbf{Y}] \text{ and } \hat{\theta}^{\text{BME}} = \underset{\theta}{\operatorname{argmax}}\, P(\theta \mid \mathbf{Y}), \tag{2}$$

where $P(\theta \mid \mathbf{Y}) \propto p(\theta) \prod_j P(Y_j = y_j \mid \theta, \hat{\boldsymbol{\zeta}}_j)$ with $p(\theta)$ being the prior distribution (e.g., Kim & Nicewander, 1993).

We aim at refining the θ estimator with a procedure that makes use of process data. Since action sequences are in a non-standard format, instead of working directly with $\boldsymbol{S}_j$, we work with the K-dimensional numerical features extracted from $\boldsymbol{S}_j$, denoted $\mathbf{X}_j = (X_{j1}, \ldots, X_{jk}) \in \mathbb{R}^K$. There are no restrictions on the feature extraction method except that the produced features $\mathbf{X}_j$ must preserve the full information on the final response Y_j, in other words, $\sigma(Y_j) \subseteq \sigma(\mathbf{X}_j)$, where $\sigma(\cdot)$ denotes the σ-algebra generated by the random variable. Intuitively, this requires the extracted features to preserve full information about the final score, such that the latter can be perfectly predicted by the process data features. Feature extraction methods such as n-gram language modeling (e.g., He & von Davier, 2016; Qiao & Jiao, 2018), multidimensional scaling (MDS; Tang

et al., 2021a), and recurrent neural network-based sequence-to-sequence autoencoders (Tang et al., 2021b), which have documented performance in terms of near-perfect final response prediction, can be applied in practice. Even if the process does not contain complete final outcome information, the final outcome can always be added as an additional dimension of the process features, to guarantee $\sigma(Y_j) \subseteq \sigma(\mathbf{X}_j)$.

1.2. Procedure

Let $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_J)$ denote the process features from all J items and $\mathbf{X}_{-j}$ the process features from the $J-1$ items excluding item j. Denote the latent trait estimate based on all J final responses by $\hat{\theta}_{\mathbf{Y}}$, which can be obtained through the estimators in Eqs. (1) or (2). Further, let $\hat{\theta}_{Y_j}$ be the estimator derived from a single response outcome Y_j, for instance, using the EAP estimator in (2).

The final response-based trait estimator, $\hat{\theta}_{\mathbf{Y}}$, can be refined by the following procedures that incorporate process features. The new estimator is denoted by $\hat{\theta}_{\mathbf{X}}$.

Procedure 1. (Construction of process-incorporated estimator)

1. *For each* $j = 1, \ldots, J$:
   - *Regress* $\hat{\theta}_{Y_j}$ *on* $\mathbf{X}_{-j}$ *to obtain* $T_{\mathbf{X}_{-j}} = E[\hat{\theta}_{Y_j}|\mathbf{X}_{-j}]$.
   - *Regress* $\hat{\theta}_{\mathbf{Y}}$ *on* $T_{\mathbf{X}_{-j}}$ *and* Y_j *to obtain* $\hat{\theta}_{\mathbf{X}_{-j}} = E[\hat{\theta}_{\mathbf{Y}}|T_{\mathbf{X}_{-j}}, Y_j]$.
2. *Compute the overall process-incorporated estimator,* $\hat{\theta}_{\mathbf{X}} = \frac{1}{J}\sum_{j=1}^{J} \hat{\theta}_{\mathbf{X}_{-j}}$.

In practice, the explicit distributions of $\hat{\theta}_{Y_j}|\mathbf{X}_{-j}$ and $\hat{\theta}_{\mathbf{Y}}|T_{\mathbf{X}_{-j}}, Y_j$ are unknown. The two conditional expectations, $E[\hat{\theta}_{Y_j}|\mathbf{X}_{-j}]$ and $E[\hat{\theta}_{\mathbf{Y}}|T_{\mathbf{X}_{-j}}, Y_j]$ in Procedure 1 are approximated on finite samples using generalized linear models. Alternatively, deep neural networks may be considered to capture nonlinear relationships. Although J regressions are required for both steps, the implementation may be paralleled to make it computationally efficient.

For illustration, consider a test of J binary items administered to N respondents. For respondent i and item j, let $\boldsymbol{s}_{ij}$ and $Y_{ij} \in \{0, 1\}$ denote the response process and the response outcome, respectively. Suppose that the response outcomes follow a two-parameter logistic model (2PL; Birnbaum, 1968),

$$\text{logit}\left(P\left(Y_{ij} = 1|\theta_i\right)\right) = a_j(\theta_i - b_j). \tag{3}$$

The following steps provide a roadmap to implement Procedure 1.

(1) IRT parameter estimation: Fit the 2PL model on the binary outcomes $\{Y_{ij} : i = 1, \ldots, N, j = 1, \ldots, J\}$ to obtain the item parameter estimates $\{\hat{\boldsymbol{\xi}}_j = (\hat{a}_j, \hat{b}_j) : j = 1, \ldots, J\}$ using marginal maximum likelihood estimation.

(2) Process feature extraction: For each item j, extract features $\mathbf{X}_{1j}, \ldots, \mathbf{X}_{Nj}$ from the problem-solving processes $\boldsymbol{S}_{1j}, \ldots, \boldsymbol{S}_{Nj}$. The MDS method (Tang et al., 2021a) or the action sequence autoencoder (Tang et al., 2021b) in the `ProcData` R package (Tang et al., 2020) can be used for this step.

(3) Response-based (baseline) latent trait estimation: One can choose from the commonly used estimators described in (1) or (2). For each respondent i and item j, get the single-item trait estimate $\hat{\theta}_{i,Y_j}$ based on Y_{ij} and $\hat{\boldsymbol{\xi}}_j$. Additionally, based on examinee i's final responses to all J items, $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iJ})$, and $\hat{\boldsymbol{\xi}} = (\hat{\boldsymbol{\xi}}_1, \ldots, \hat{\boldsymbol{\xi}}_J)$, obtain $\hat{\theta}_{i,\mathbf{Y}}$.

(4) First conditional expectations: For each j, fit a regression $\hat{\theta}_{Y_j} \sim \mathbf{X}_{-j}$ based on $\{(\mathbf{X}_{i(-j)}, \hat{\theta}_{i,Y_j}) : i = 1, \ldots, N\}$ to approximate $E[\hat{\theta}_{Y_j}|\mathbf{X}_{-j}]$ and calculate the fitted values $\{T_{i,\mathbf{X}_{-j}} : i = 1, \ldots, N\}$. For example, one can use ridge regression (Hoerl & Kennard, 1970) with shrinkage parameter selected by cross-validation.

(5) Second conditional expectations: For each j, fit another regression $\hat{\theta}_{\mathbf{Y}} \sim (T_{\mathbf{X}_{-j}}, Y_j)$. One simple choice is ordinary least squares with $(1, T_{\mathbf{X}_{-j}}, Y_j, T_{\mathbf{X}_{-j}} Y_j)$ as predictors, where $T_{\mathbf{X}_{-j}} Y_j$ is the interaction term.

(6) Averaging step: The average of the fitted values $\{\hat{\theta}_{i,\mathbf{X}_{-j}} : j = 1, \ldots, J\}$ in step (5) is the final process-incorporated trait estimate, $\hat{\theta}_{i,\mathbf{X}}$, for respondent i.

1.3. Theoretical Analysis

In this subsection, we provide theoretical justifications for this procedure. In particular, we elaborate a set of assumptions under which the process data score improves latent trait estimation in terms of mean squared error. The first assumption requires the conditional expectation of $\hat{\theta}_{Y_j}$ given θ to be monotonically increasing. This assumption is satisfied by well-designed cognitive items and latent trait estimators.

A1. (Monotonicity assumption) $m_j(\theta) = E\left[\hat{\theta}_{Y_j}|\theta\right]$ is monotone in θ and has a finite second moment.

Secondly, we assume that the response outcome of item j is correlated with an individual's behaviors on other items only through the measured trait θ and not through other latent or observed traits. Since the process features can include rich information other than the measured trait, this local independence assumption requires Y_j to be "good", in the sense that no construct-irrelevant, persistent traits affect final performance. In other words, measurement error comes from sources of random, instead of systematic error (AERA, APA, and NCME, 2014). For example, the process features $\mathbf{X}_{-j}$ may reflect a respondent's computer usage habits, such as whether they tend to double- or single-click on buttons. However, the final score, Y_j, shall not differentiate individuals with different clicking habits, as long as they have the same level of θ. We do allow Y_j to be very "rough" measurements, in other words, the measurement error can be large, as long as it is due to random instead of systematic variations. Similarly, $\hat{\theta}_{Y_j}$ can be biased and can have a large standard error, as long as the monotonicity assumption (A1) is satisfied.

A2. (Local independence assumption) Conditioning on the latent trait θ, Y_j and $\mathbf{X}_{-j}$ are independent.

Finally, we consider the distribution of the process features, $\mathbf{X}_{-j}$, given the measured trait. Note that the problem-solving process can depend on traits other than θ. These unobserved, construct-irrelevant traits are assumed random and integrated out from the probability density, thus resulting in the conditional density of $\mathbf{X}_{-j}$ given θ. We impose the usual exponential family assumption on process features for technical development. This is equivalent to assuming the existence of a unidimensional sufficient statistic of θ independent of sample size (Lehmann & Romano, 2005). The natural parameter $\eta_j(\theta)$ is assumed to be monotone so that there is no identifiability issue for θ.

A3. (Exponential family assumption) The probability density function for features $\mathbf{X}_{-j}$ for each j takes the following form:

$$f(\mathbf{X}_{-j}|\theta) = \exp\left\{\eta_j(\theta)\, T_j(\mathbf{X}_{-j}) - A_j(\theta)\right\} h_j(\mathbf{X}_{-j}), \tag{4}$$

where $T_j(\mathbf{X}_{-j})$ is a sufficient statistic for θ and the natural parameter $\eta_j(\theta)$ is monotone in θ with a finite second moment.

Theorem 1 shows that the first step of our proposed procedure summarizes the process data features to sufficient statistics.

Theorem 1. *Under Assumptions A1–A3, $T_{\mathbf{X}_{-j}}$ is a sufficient statistic of $\mathbf{X}_{-j}$ for θ.*

Based on the sufficiency of $T_{\mathbf{X}_{-j}}$, we can further show that $\hat{\theta}_{\mathbf{X}}$ reduces the MSE of $\hat{\theta}_{\mathbf{Y}}$, as stated in Theorem 2. The proof of this result uses the Rao–Blackwell theorem (Blackwell, 1947; Casella & Berger, 2002) and also shows that every $\hat{\theta}_{\mathbf{X}_{-j}}$ produced by step 2 of the procedure removes conditional variance and improves $\hat{\theta}_{\mathbf{Y}}$ in terms of MSE. The proofs of Theorem 1 and Theorem 2 are provided in Appendix.

Theorem 2. *If assumptions A1-A3 hold for all J items, then*

$$E[(\hat{\theta}_{\mathbf{X}} - \theta)^2|\theta] \leq E[(\hat{\theta}_{\mathbf{Y}} - \theta)^2|\theta] \quad \textit{for every } \theta. \tag{5}$$

The MSE reduction by incorporating process features implies an increase in statistical efficiency for estimating θ. Putting Theorem 2 in the psychometric context, the MSE reduction of θ estimator translates to the reduction of *conditional standard error of measurement* at all possible proficiency levels, therefore, higher measurement reliability. In practice, this allows one to derive more reliable scores (i.e., latent trait estimates) for individuals in assessments. Alternatively, by using the procedure to incorporate process data, one is able to achieve comparable measurement precision to traditional outcome-based scoring with fewer items.

Here, we explain the rationale behind the assumptions and expected consequences of assumption violation: Assumption A1 is the most important among all. It provides a crucial guide to extracting the relevant part of the process information on the measured trait. This also suggests that the final-response-based IRT model yields a valid estimator with random noise, written as $\hat{\theta}_{Y_j} = m_j(\theta) + \varepsilon_j$. Assumption A2 ensures that ε_j is not predictable by the process data of other items, and furthermore, ε_j has zero mean across the population. Thus, if A2 is not valid, we may expect a certain amount of bias introduced to the process estimator. A3 provides a technical framework for us to discuss efficiency. We choose the natural exponential family because it is the first-order approximation of a large class of parametric families.

2. Simulations

Simulation studies were conducted to compare outcome- and process-incorporated estimators of the latent trait.

2.1. Experiment Settings

We generated respondents' latent trait $\theta_1, \ldots, \theta_N$ independently from the standard normal distribution. Given examinee i's underlying true θ_i, the response process and the final response to each item were then generated, both dependent on the latent trait θ_i. Specifically, for respondent i and item j, the response outcome Y_{ij} followed a Rasch model (Rasch, 1960),

$$\text{logit}\left(P\left(Y_{ij} = 1|\theta_i\right)\right) = \theta_i - b_j. \tag{6}$$

To generate the problem-solving process, we considered a Markov model and an action set of 26 English letters. Each letter represents 1 of 26 possible actions recorded at each time point. The

probability transition matrix was distinct for each respondent-item pair and denoted as $\mathbf{P}^{(ij)} = (p_{kl}^{(ij)})_{1\le k,l\le M}$ for the ith respondent and the jth item. Given the probability transition matrix, we generated an action sequence starting from "A", where the subsequent actions were sampled according to $\mathbf{P}^{(ij)}$ until the final state "Z" appears. Excluding the column for "A" and the row for "Z", the upper right $(M-1)\times(M-1)$ submatrix of $\mathbf{P}^{(ij)}$ was computed according to

$$p_{kl}^{(ij)} = \frac{\exp(\theta_i u_{kl}^{(j)})}{\sum_{r=1}^{M-1}\exp(\theta_i u_{kr}^{(j)})}, \tag{7}$$

where $(u_{kl}^{(j)})_{1\le k,l\le M-1}$ were generated independently from Uniform$(-10, 10)$ for each item. Each item's action sequences hence depended on the underlying θ through the transition probabilities between actions. Note that, by the current simulation design, the final response, generated separately, was not a function of the process and hence not perfectly predictable from process. To ensure the perfect predictability of response from process features on item j, we included the final response as an additional dimension for the process features in subsequent analysis.

Two experiments were devised to evaluate the effect of sample size (N) and test length (J). Experiment I considers four different sample sizes: $N = 200, 500, 1000$ and 2000. The number of items J was fixed to three with difficulty parameters $b_1 = 0, b_2 = 1, b_3 = -1$ in (6). Each condition of N was replicated 100 times. Experiment II considers different test lengths. We considered a maximum of 20 items and generated the difficulty parameter b_j from Uniform$(-1, 1)$ for each item. Starting from two items, we added one more observed item for estimation in each step until all 20 items were included. The sample size N was fixed at 2000.

MDS features were extracted from the response process on each item, with latent dimension K chosen by five-fold cross-validation from candidate values $10, 20, \ldots, 50$. To guarantee perfect predictability of Y_j by $\mathbf{X}_j$, the final response to each item was added as an additional dimension to the process features. Following the illustration in Sect. 1.2, we first estimated the item parameters b_1, b_2, b_3 by marginal maximum likelihood estimation. Then, we used response outcomes to calculate the baseline EAP estimator, $\hat{\theta}_{\mathbf{Y}}$, as well as the single-item response-based EAP estimators, $\hat{\theta}_{Y_1}, \ldots, \hat{\theta}_{Y_J}$. Note that by using the EAP, we minimize the posterior MSE. Ridge regression (Tikhonov & Arsenin, 1977) was used for the first conditional expectation, $E[\hat{\theta}_{Y_j}|\mathbf{X}_{-j}]$, and the shrinkage parameter was tuned to minimize the deviance in fivefold cross-validation. For the second conditional expectation, $E[\hat{\theta}_{\mathbf{Y}}|T_{\mathbf{X}_{-j}}, Y_j]$, we regressed $\hat{\theta}_{\mathbf{Y}}$ on $(1, T_{\mathbf{X}_{-j}}, Y_j, T_{\mathbf{X}_{-j}}Y_j)$ by ordinary least squares.

2.2. *Results*

The estimators $\hat{\theta}_{\mathbf{Y}}$ and $\hat{\theta}_{\mathbf{X}}$ were evaluated by two criteria, MSE and Kendall's rank correlation (τ; Kendall, 1938). The MSE of an estimator $\hat{\theta}$ was calculated by

$$\mathrm{MSE}(\hat{\theta}) = \frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i)^2, \tag{8}$$

where $\hat{\theta}_i$ is the estimate for the i-th respondent, and θ_i is the true latent trait of respondent i. The Kendall's τ between estimated and true θ can also be calculated for both estimators. In contrast to MSE, Kendall's τ considers the extent to which the estimated ranking aligns with the true ranking of latent trait, which is the interest of norm-referenced tests.

As shown in Fig. 1, the process-incorporated estimator (x-axis) outperformed the outcome-based estimator (y-axis) in terms of both MSE and Kendall's τ. In each subplot, the 100 points

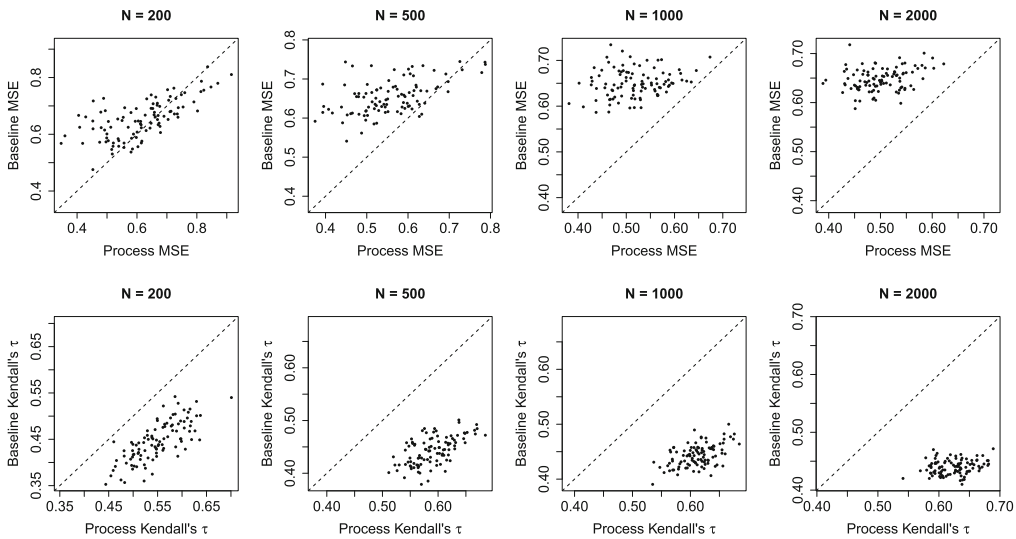

FIGURE 1.
Response-based (baseline) and process-incorporated estimators' agreement with true θ across different sample sizes.

correspond to results from 100 replications. For smaller sample sizes of $N = 200$ or 500, the MSE of the process-incorporated estimator was higher than that of the response-based estimator in some replications. As the sample size increased, the proposed procedure consistently achieved smaller MSE. The Kendall's τ of the process-incorporated estimator was consistently higher than that produced from the baseline estimator across sample sizes and replications, and the improvement became more substantial when N increased. As shown in the subplot for $N = 2000$, with three items, τ could increase from around 0.45 to over 0.60 after applying the procedure to incorporate process information. Figure 2 takes a closer look at MSE by dividing respondents into 10 groups based on their true latent trait, with $N = 2000$. Group 1 has the lowest θ and group 10 has the highest θ. Within each group, we calculated MSE for baseline and process-incorporated estimators over the 100 simulated data sets. The box plots show that for groups 2 to 9, the process-incorporated estimator substantially reduced MSE. For examinees with extreme true θs (groups 1 and 10), both estimators had larger errors.

Figure 3 displays the results of experiment II where test length was considered. In the left panel, the MSE of the process-incorporated estimator (red line) remained below that of the outcome-based estimator (black line) as the number of items increased from 2 to 20. The proposed procedure reduced the MSE by over a half. The improvement in Kendall's τ was also consistent across different test lengths as shown in the right panel. For instance, when $J = 7$, Kendall's τ rose from 0.5 to 0.7 after the refinement.

3. Empirical Example: PIAAC PSTRE

The proposed approach for score refinement was further applied to the data collected from the PSTRE assessment from the 2012 PIAAC survey. The empirical analyses were guided by two overarching objectives. First, the performance of the response- and process-incorporated latent trait estimators was compared, similar to the simulation studies. Second, because response-based and process-incorporated estimators are expected to produce different latent ability estimates

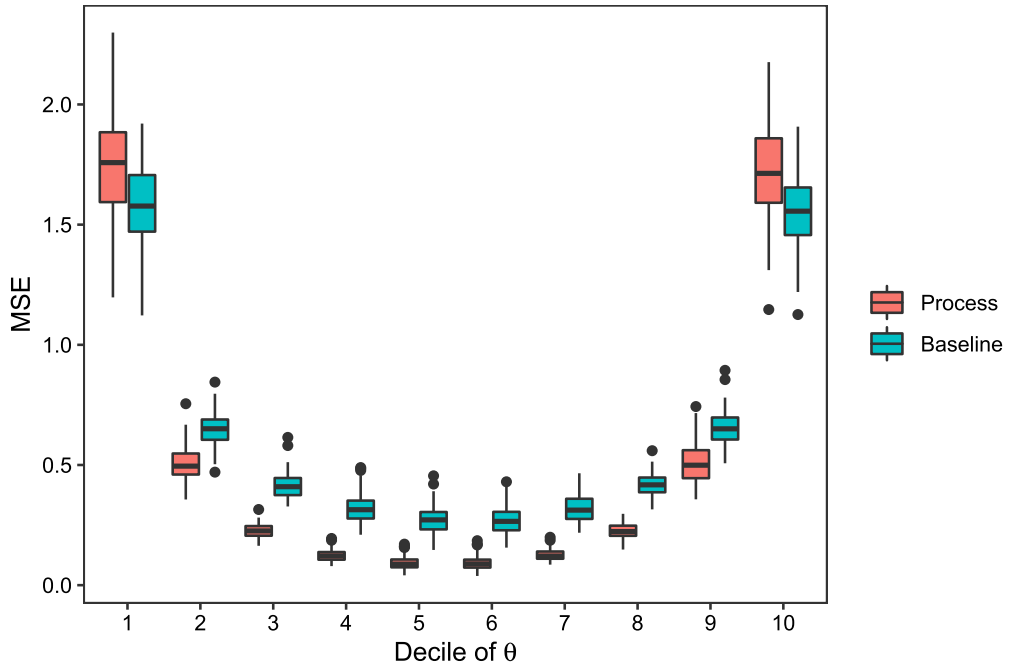

FIGURE 2.
Box plots of MSE in different strata of true θ ($N = 2000$).

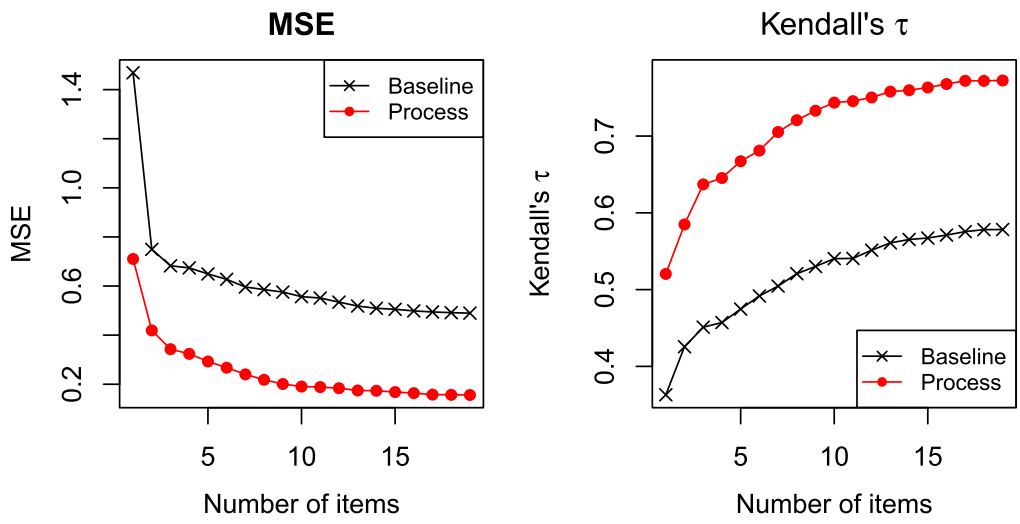

FIGURE 3.
Simulation results of experiment II.

for the same examinee, we further examined the problem-solving patterns associated with large discrepancies in response- and process-incorporated $\hat{\theta}$s. In the following subsections, a description of the PIAAC PSTRE data is first provided, followed by the methods and findings from the empirical analyses.

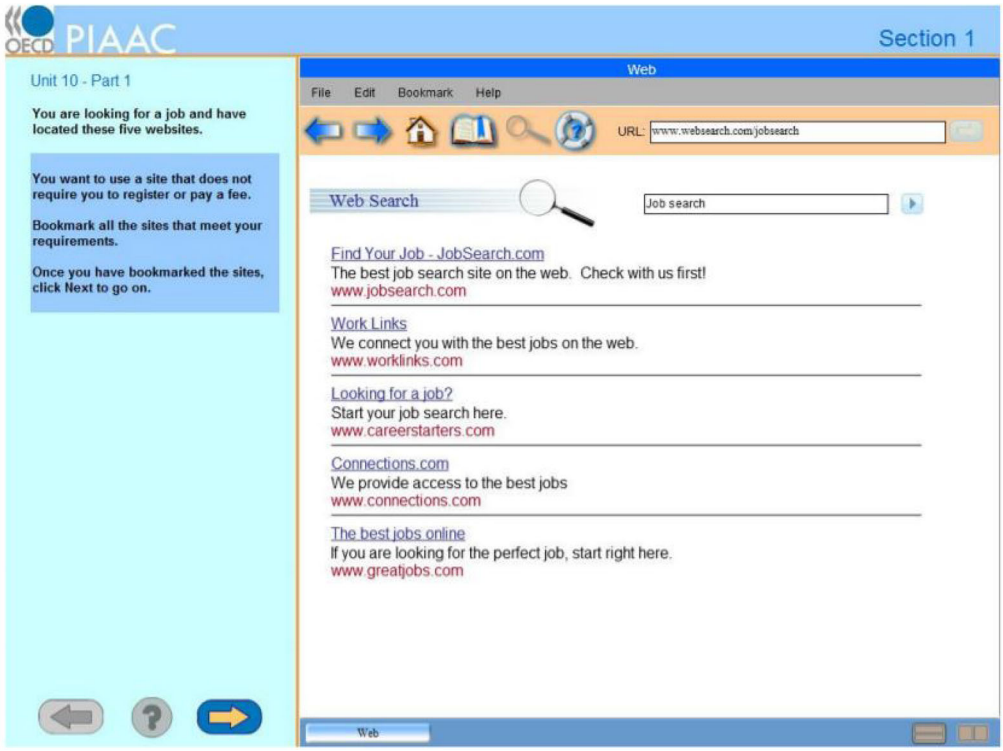

FIGURE 4.
Home page of the PSTRE sample item. Reprinted from OECD Sample Questions and Questionnaire (https://www.oecd.org/skills/piaac/samplequestionsandquestionnaire.htm).

3.1. The PIAAC PSTRE Data

Carried out by the Organization for Economic Co-operation and Development (OECD), the PIAAC (e.g., Schleicher, 2008) is an international survey of the cognitive and workplace skills of working-age individuals around the world. The first cycle of the PIAAC survey in 2012 assessed three cognitive skills, namely literacy, numeracy, and PSTRE, on participants from 24 countries and regions with ages between 16 and 65 years. In addition to the three cognitive assessments, the participants were further surveyed on their demographic background and other information related to their occupation and education.

The current study focuses on the PSTRE assessment, where individuals were administered a series of computer-based interactive items. PSTRE ability refers to the ability to use digital technology, communication tools, and the internet to obtain and evaluate information, communicate with others, and perform practical tasks (OECD, 2012). Successful completion of the PSTRE tasks requires both problem-solving skills and familiarity with digital environments. The test environment of each item resembled commonly seen informational and communicative technology (ICT) platforms, such as e-mail clients, web browsers, and spreadsheets. Test takers were asked to complete specific tasks in these interactive environments. Individuals' entire log of interactions with each item was recorded as log data. In addition, based on the extent of task completion, polytomous final scores were derived for each item.

A sample item that resembles PSTRE tasks is shown in Fig. 4. Respondents can read the task instructions on the left side and work on the task in the simulated interactive environment on the right. This item requires respondents to identify, from the five web pages presented on the screen,

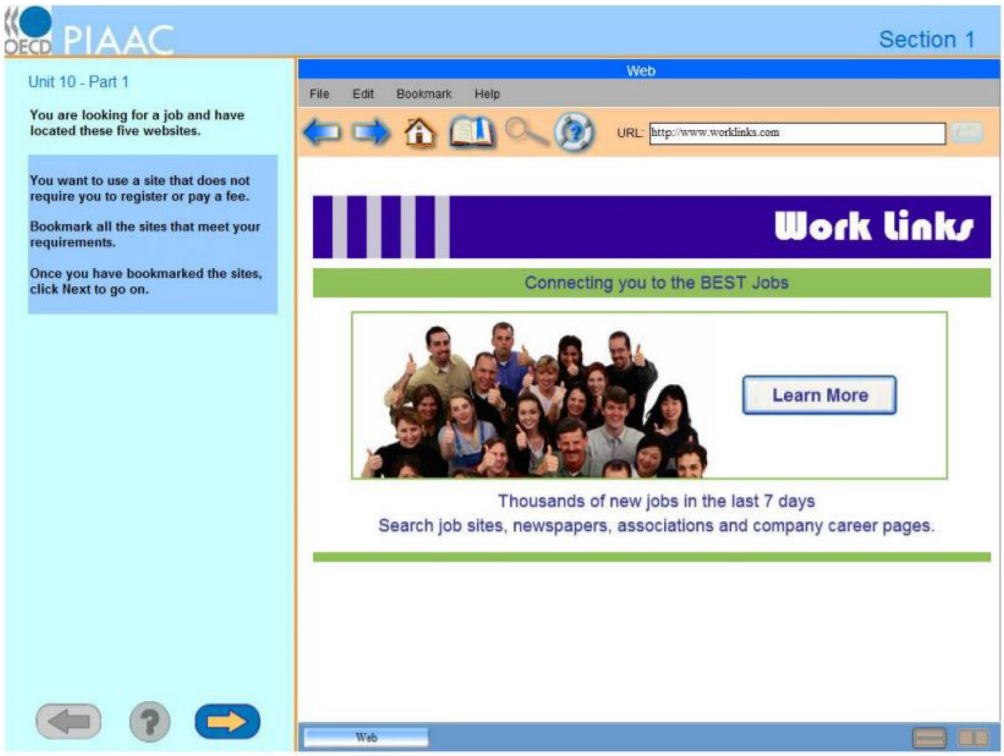

FIGURE 5.
Web page returned from clicking the second link (i.e., "Work links") on the home page. Reprinted from OECD Sample Questions and Questionnaire (https://www.oecd.org/skills/piaac/samplequestionsandquestionnaire.htm).

all pages that do not require registration or fees and to bookmark them. By clicking on each link, they will be redirected to the corresponding website where they can learn more. For example, clicking "Work Links" directs them to Fig. 5, and further clicking on "Learn More" directs them to the page in Fig. 6. Once having finished working on the task, a test taker can click on the right arrow ("Next") on the bottom-left. A pop-up window will ask them to confirm their decision by clicking "OK" or to return to the question by clicking "Cancel". A respondent who clicked on the aforementioned two links, bookmarked the page using the toolbar icon, and moved on to the next question will have the recorded action sequence of "Start, Click_W2, Click_Learn_More, Toolbar_Bookmark, Next, Next_OK".

The computer-based version of the 2012 PIAAC survey randomly assigned each respondent with two blocks of cognitive items, where each block consisted of a fixed set of items that assessed either literacy, numeracy, or PSTRE proficiency. The current study uses the PSTRE response and process data of individuals from five countries and regions, including the UK (England and Northern Ireland), Ireland, Japan, the Netherlands, and the USA, who were assigned to PSTRE for both blocks. The five countries and regions were relatively similar in performance distribution on the PIAAC PSTRE assessment. Each PSTRE block consisted of 7 items, and thus, the two blocks total to 14 items. Note that a recorded action sequence of "Start, Next, Next_OK" indicates that the test taker did not perform any action on the item and moved on to the next question. This type of behavior can be regarded as omission and is distinguished from either credited or uncredited responses. The current study excluded individuals who omitted any of the 14 items, resulting in a total of 2304 test takers who responded to all 14 PSTRE items. For each item, the action sequence of

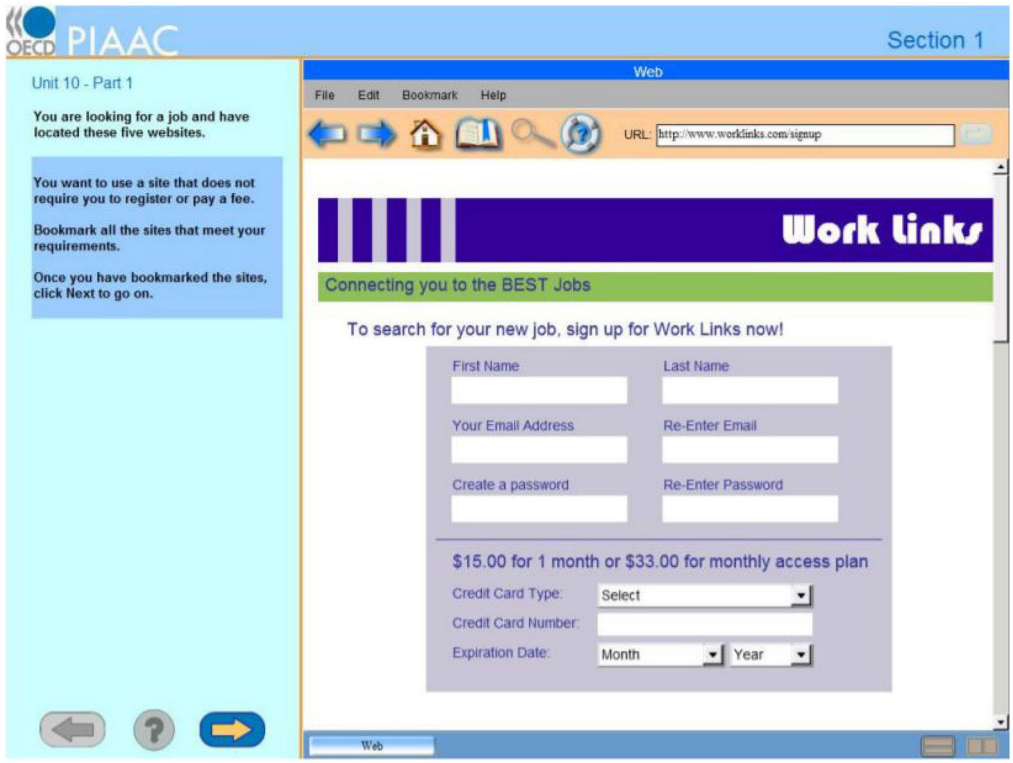

FIGURE 6.
Web page returned from clicking "Learn More" on the "Work links" website. Reprinted from OECD Sample Questions and Questionnaire (https://www.oecd.org/skills/piaac/samplequestionsandquestionnaire.htm).

each test taker was recorded, and a polytomous final score calculated based on predefined scoring rubrics was available. PIAAC uses the final scores together with other demographic covariates to obtain plausible values of individuals' proficiency on PSTRE. Table 1 presents descriptive information of the 14 PSTRE items, including the task names and the descriptive statistics of the final scores and action sequences.

3.2. Overall Performance in Latent Proficiency Estimation

3.2.1. Evaluation Methods With empirical data, respondents' true θs were unknown. The two proficiency estimators were instead compared on their agreement with performance on a separate set of items designed to measure the same trait. Specifically, the 14 PSTRE items were split into two sets of 7 items. One set of 7 items, denoted the scoring set ($\mathcal{B}_s$), was used to obtain the response- and the process-incorporated estimators ($\hat{\theta}_{\mathbf{Y}}^{(s)}$ and $\hat{\theta}_{\mathbf{X}}^{(s)}$). A separate latent trait estimate, $\hat{\theta}_{\mathbf{Y}}^{(r)}$, can be obtained from the final responses to the remaining 7 items, denoted the reference set ($\mathcal{B}_r$). Any trait estimate obtained from the scoring set does not use reference set response information, and $\theta_Y^{(r)}$ serves as an external criterion for evaluating $\hat{\theta}_{\mathbf{Y}}^{(s)}$ and $\hat{\theta}_{\mathbf{X}}^{(s)}$. The 14 items can be partitioned into scoring and reference sets in $\binom{14}{7}$ ways. We randomly chose 50 possible partitions and evaluated the agreement on each partition. On a remark, the original two blocks of 7 items were thoughtfully designed to be parallel. Here, the forms were scrambled for method evaluation, where the resulting partitions may no longer be comparable in specific content coverage, difficulty, and other item characteristics, despite measuring a common unidimensional trait of PSTRE.

TABLE 1.
Descriptive information of the 14 PIAAC PSTRE items.

Item ID	Task name	Final score		Action types	Sequence length		
		Score levels	Median		Min	Max	Median
U01a	Party invitations	4	3	40	4	90	17
U01b	Party invitations	2	1	47	4	132	29
U02	Meeting room	4	1	95	4	153	35
U03a	CD tally	2	1	67	4	51	9
U04a	Class attendance	4	0	615	4	304	49
U06a	Sprained ankle	2	0	30	4	57	10
U06b	Sprained ankle	2	1	26	4	51	18
U07	Book order	2	1	40	4	79	24
U11b	Locate email	4	2	122	4	256	22
U16	Reply all	2	1	359	4	267	34
U19a	Club membership	2	1	75	4	356	19
U19b	Club membership	3	2	244	4	396	18
U21	Tickets	2	1	124	4	77	22
U23	Lamp return	4	3	133	4	139	25

Descriptive statistics calculated based on the 2304 participants without omission; Score levels: number of ordinal response categories; Action types: the number of possible actions in the log data; Sequence length: the number of actions recorded in a test taker's process data.

Similar to the simulation study, the mean-squared deviation (MSE) with respect to $\hat{\theta}_{\mathbf{Y}}^{(r)}$,

$$\text{MSE}(\hat{\theta}^{(s)}) = \frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}^{(s)} - \hat{\theta}_{\mathbf{Y}}^{(r)})^2, \tag{9}$$

and the Kendall's τ with $\hat{\theta}_{\mathbf{Y}}^{(r)}$ can be computed for each estimator produced from the scoring set. Unlike the true θ, $\hat{\theta}_{\mathbf{Y}}^{(r)}$ is estimated based on final responses to only 7 items and contains measurement error. The correlation between $\hat{\theta}^{(s)}$ and $\hat{\theta}_{\mathbf{Y}}^{(r)}$ is hence attenuated by the reliability of $\hat{\theta}_{\mathbf{Y}}^{(r)}$, and the MSE of $\hat{\theta}^{(s)}$ with respect to $\hat{\theta}_{\mathbf{Y}}^{(r)}$ is expected to deviate from the MSE of $\hat{\theta}^{(s)}$ with respect to true θ. Rather than interpreting the two evaluation metrics as the recovery of true proficiency, they can instead be regarded as the split-half ($\mathcal{B}_s$ and $\mathcal{B}_r$) agreement of latent trait estimates, or, alternatively, as the strength of association between $\hat{\theta}^{(s)}$ and performance on similar tasks ($\hat{\theta}_{\mathbf{Y}}^{(r)}$). Lower MSE and higher Kendall's τ hence suggest higher reliability.

On the scoring set, the response- and process-incorporated estimators were obtained following similar procedures as in the simulation studies. Specifically, the dimension of the process features was selected for each item by cross-validation, and the final score to each question was added as an additional process feature to guarantee perfect predictability. To evaluate performance under different test lengths, similar to experiment II, the number of items in $\mathcal{B}_s$ used for scoring ranged from 2 to 7. Specifically, when only two items were used for scoring, it was assumed that examinees' processes and scores were observed only on the first two items. Subsequent items in the scoring set were added one by one until all 7 items were used. Because the final responses were polytomous, the generalized partial credit model (Muraki, 1992) was used to calibrate the item parameters and to obtain the response-based θ estimates. Additionally, for the second conditional expectation in Procedure 1, $E(\hat{\theta}_{\mathbf{Y}} \mid T_{\mathbf{X}_{-j}}, Y_j)$, the polytomous final response Y_j was dummy-coded as regressors in the second regression.

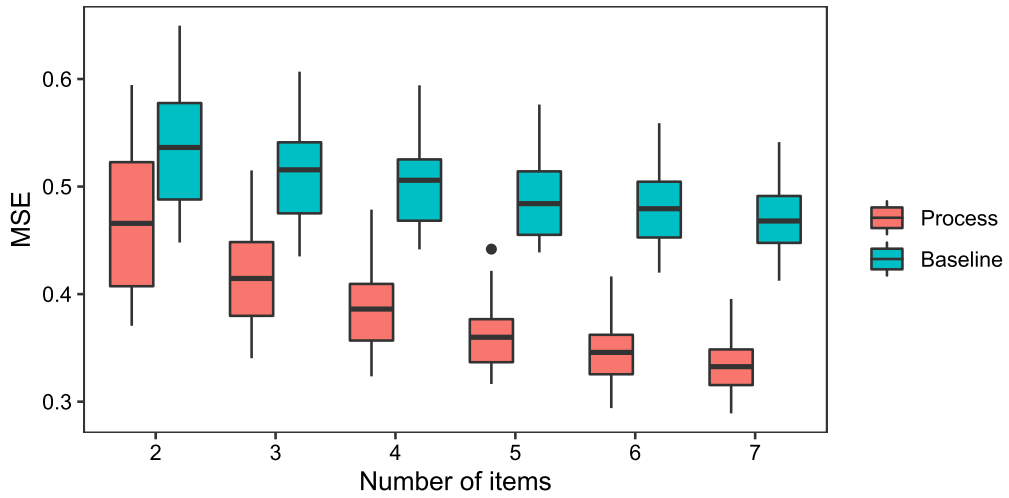

FIGURE 7.
Response-based (baseline) and process-incorporated estimators' MSE with respect to reference-set ability estimate on the PIAAC data. x-axis is the number of items used for computing $\hat{\theta}_{\mathbf{X}}$ and $\hat{\theta}_{\mathbf{Y}}$ in the scoring set. Box plots represent the distribution of MSE across the 50 partitions of $\mathcal{B}_s$ and $\mathcal{B}_r$.

3.2.2. Results Figure 7 compares the MSE of the response- and process-incorporated trait estimators. Results are presented for different test lengths, ranging from 2 to 7 items, in the scoring set. The green boxes correspond to the response-based (baseline) $\hat{\theta}^{(s)}$, and the red boxes correspond to the process-incorporated $\hat{\theta}^{(s)}$. Each box plot in Fig. 7 represents the distribution of the MSE across the 50 randomly sampled partitions of the 14 items into scoring and reference sets. One can observe that for all test lengths, the process-incorporated latent trait estimator consistently demonstrated smaller MSE, indicating higher agreement with the performance on an external set of similar tasks (i.e., the reference set). In particular, with two items, the process-incorporated estimator achieved comparable median MSE as the response-based estimator using five items. With four or more items, the process-incorporated $\hat{\theta}$ consistently achieved similar, if not lower, MSE than the response-based estimator using all 7 items.

The box plots for the Kendall's τ of the two types of estimators are presented in Fig. 8. The correlations with reference set performance were consistently larger using process-incorporated scoring for all test lengths, suggesting that the rankings of latent ability estimates generated based on the problem-solving processes were more similar to the rankings on reference set performance. Again, scores based on processes required fewer items to achieve a given level of agreement. For instance, with 4 items, the process-incorporated estimator achieved similar or higher Kendall's τ when compared to the response-based estimator with all 7 items. Attenuated by the reliability of the reference set θ estimate, the absolute Kendall's τ were mostly below .5. When compared to the true θ, however, one would expect the correlation to be higher.

3.3. Performance by Degree of Process- and Response-Based Score Discrepancy

3.3.1. Evaluation Methods The comparisons above focused on the overall agreement of each estimator with reference set performance. One may also be interested in how the two methods perform for different types of examinees. In particular, it is worth evaluating the relative performance of the two estimators when they disagree on an examinee's latent proficiency ranking. On each of the 50 partitions, we computed the process-incorporated and response-based estimators using all 7 items on the scoring set. We further regressed the response-based $\hat{\theta}_{\mathbf{Y}}^{(s)}$ on the

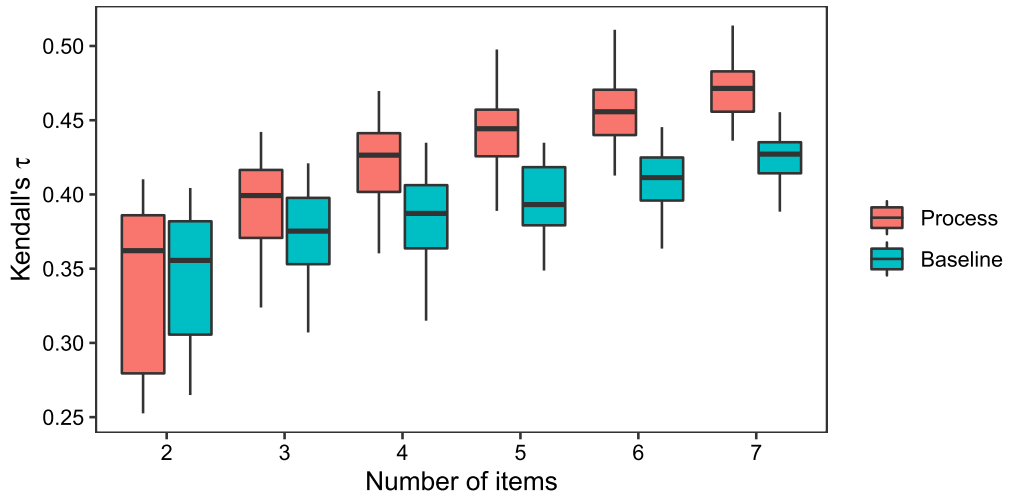

FIGURE 8.
Distribution of response-based (baseline) and process-incorporated Kendall's τ with reference set performance across 50 partitions on the PIAAC data.

process-incorporated $\hat{\theta}_{\mathbf{X}}^{(s)}$ using OLS and computed each individual's Studentized residual for the regression. Individuals were then binned into 10 groups based on their deciles of the Studentized residuals. The deciles of the Studentized residuals reflect the relative discrepancies in performance ranking based on the two trait estimators: For individuals in the first decile, their performance rankings based on process were much lower than that based on responses. Individuals in the 10th decile, on the other hand, were ranked much higher based on responses than based on process. Individuals closer to the middle (4th–6th decile) received similar rankings based on process and responses. The MSEs of the two trait estimators with respect to $\hat{\theta}_{\mathbf{Y}}^{(r)}$ within each decile were then computed.

3.3.2. Results The box plots of the MSEs with respect to reference set performance $\hat{\theta}_{\mathbf{Y}}^{(r)}$ across the 50 partitions, separated by residual deciles, are shown in Fig. 9. When the two scores agree on individuals' rankings, the MSEs of $\hat{\theta}_{\mathbf{Y}}^{(s)}$ and $\hat{\theta}_{\mathbf{X}}^{(s)}$ were similar. However, as we move towards the two ends where the two estimators started to disagree, the MSEs of the process-incorporated estimator were remarkably lower than that of the response-based estimator. Intuitively, the two estimators can be thought of as two judges, one judging individuals' performance considering the problem-solving processes, and the other judging solely based on the final outcome. When the two judges disagree, the process-incorporated estimator consistently better predicted performance on similar tasks.

3.4. Empirical Interpretations of Process- and Response-Based Score Discrepancy

3.4.1. Methods The results above suggest that the proposed process-incorporated latent trait estimation procedures led to an increase in consistency with proficiency estimate on an external set of items, and that the improvement appeared most significant for individuals whose process-incorporated and response-based latent trait estimates disagreed most. One question worth asking is how the proposed approach scores individuals differently compared to the response-based counterpart. We explored this question by looking at the sequences of individuals whose process-incorporated and response-based latent trait estimates disagreed the most, that is, those with the highest or lowest Studentized residuals for $\hat{\theta}_{\mathbf{Y}} \sim \hat{\theta}_{\mathbf{X}}$.

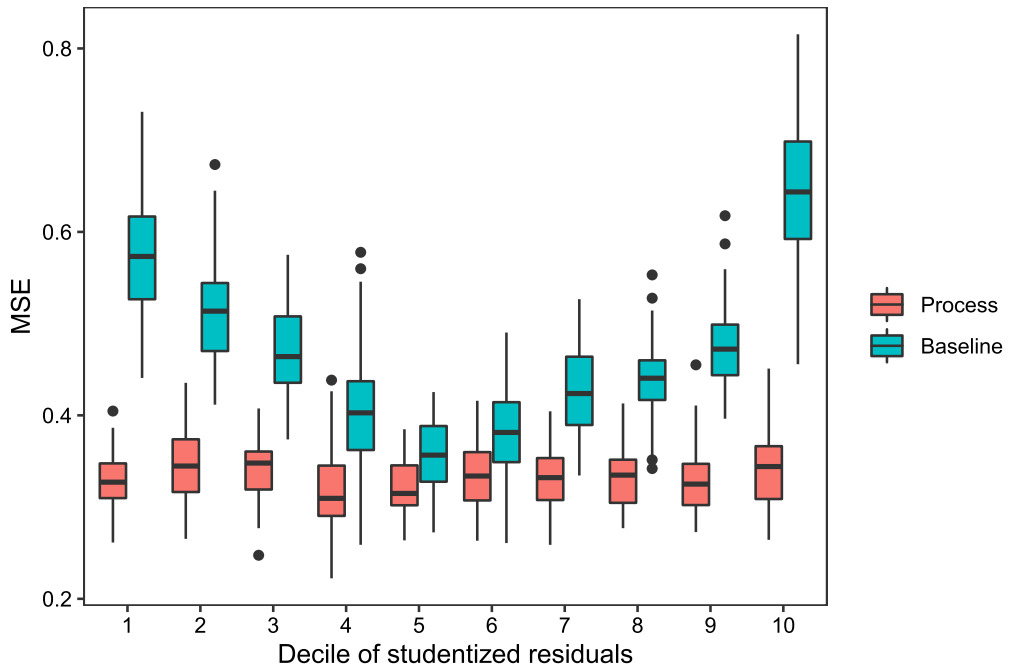

FIGURE 9.
Distribution of MSEs with reference set latent trait estimate in each residual decile. Box plots are the distributions of MSEs within each decile bin across the 50 partitions of scoring and reference sets.

This time, with the purpose of interpretation rather than performance evaluation, all 14 items were used to obtain the two trait estimates. For the individuals in the bottom and top 10 of the Studentized residuals, we visually examined their action sequences on the 14 items.

3.4.2. Results Figure 10 shows the scatter plot of each respondent's $\hat{\theta}_{\mathbf{X}}$ (x-axis) and $\hat{\theta}_{\mathbf{Y}}$ (y-axis). Blue triangles correspond to the 10 individuals with the highest Studentized residuals, when regressing $\hat{\theta}_{\mathbf{Y}}$ on $\hat{\theta}_{\mathbf{X}}$. These individuals received lower rankings based on processes than based on final responses. For most of them, certain questions were successfully solved, but with less efficient strategies: For instance, to look for the requested information from a long spreadsheet, some of these examinees visually inspected every single entry, although a much more efficient strategy is to use "Search" or "Sort" (e.g., items U03a, U19a). To reply to an email sent to a group, some of them hand-typed the long list of email recipients, when they could simply press "Reply to all" or copy-and-paste (e.g., item U16). Aside from inefficient strategy usage, a few examinees also performed a large number of redundant steps, that is, actions that were not required for successful task completion.

The red rectangles in Fig. 10, on the other hand, represent the 10 examinees with the lowest Studentized residuals. These examinees received higher ranking based on their problem-solving processes than based on final scores. Several common patterns were observed from their log data: The first was partial completion, where the examinee performed some of the key steps on a question, but, before reaching a credited response, proceeded to the next question by clicking "Next, Next_OK". An example is item U16, which required sending an email to a list of recipients containing some key information. Several of the 10 examinees created the email and filled in the correct content and recipients, but they proceeded to the next question without clicking "Send". Another common pattern was careless mistakes, where the examinee demonstrated the required

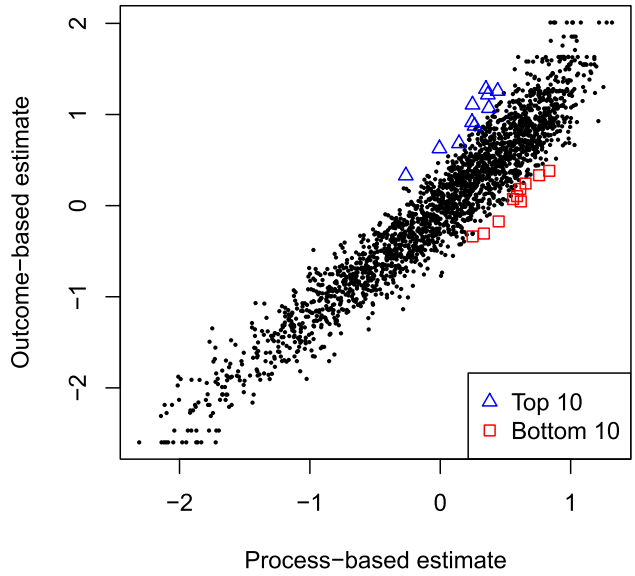

FIGURE 10.
Scatterplot of process-incorporated and response-based θ estimates with 14 items.

skills for completing the task but slipped on an item due to carelessness. For example, on item U11b, which required sorting emails in the "Saved" folder, four of the ten examinees sorted the emails in the "Inbox", the default, wrong folder. Intuitively, occasional carelessness and misinterpretation of question requirements, which lead to incorrect responses despite having the requisite skills, may be regarded as one of many sources of random measurement error. With additional information from the problem-solving processes incorporated, the proposed procedure for process-incorporated scoring appeared less impacted by such sources of measurement error.

4. Discussions

4.1. Findings and Implications

Problem-solving processes contain rich information on individual characteristics, including the measured construct. The current study introduces a method to refine outcome-based latent trait estimates using the additional information from the problem-solving processes. A Rao–Blackwellization approach was proposed for the score refinement. Aside from choosing an appropriate IRT model for the final responses, the proposed approach is relatively data-driven and does not involve prior specification of a measurement model for the problem-solving processes, requiring less subjective inputs compared to expert-defined rubrics for process-incorporated scoring.

The main theorem states that under some regularity conditions, the proposed approach can lead to MSE reduction in latent trait estimation. Results from simulation studies corroborate the theorem. An empirical study using the PIAAC PSTRE data further showed that the process-incorporated latent trait estimate tended to have higher agreement with performance on similar tasks, thus higher reliability, compared to the response-based trait estimate. In addition, in order to achieve a particular level of reliability (i.e., MSE or τ with the external set of items), far

fewer items would be required if the additional information from the problem-solving processes is exploited for scoring.

While the improvement in reliability by incorporating process information was consistent across test lengths, one particular merit of incorporating the process information into scoring lies in its use for short tests: For short tests, the final responses alone can rarely generate reliable trait estimates. Process information can thus be particularly useful for low-stakes computer-based assessment scenarios, when the administration of long tests is unrealistic or burdensome. With the additional information from problem-solving processes, the tests can be significantly shortened without compromising measurement reliability. An example is interim formative assessments during the learning process, where, after every one or few classes, instructors may want to learn how well the students have mastered the recently taught contents. Administration of a long test after every several classes can be very burdensome for the students and may interrupt the learning process. On the other hand, a relatively reliable latent ability estimate can be obtained if the problem-solving processes to a few constructed response items are available. Although computerized adaptive testing (CAT; e.g., Wainer et al., 2000) can also reduce the required test length through the adaptive selection of test items tailored to individuals' real-time proficiency estimates, the construction of a CAT usually requires a large pre-calibrated item pool with hundreds of items, which may be overly costly and hard to achieve for many smaller-scale and low-stakes assessments. The production of a process-incorporated scoring rule, on the other hand, only requires sufficient items for reliable measurement of latent proficiency and a sample size that is sufficient for item parameter calibration, process feature extraction, and training the regression models.

4.2. Practical Considerations

The proposed approach could consistently improve test reliability. However, there are a few caveats to its implementation, especially for examinations with higher stakes. First, in the empirical study, the performance of the process- and response-based latent trait estimators was evaluated using up to 7 items for scoring. The choice of up to 7 items was due to the limited number of total items available (14) and the need to set aside a large enough reference set of items used for evaluations. For an operational test, however, 7 items' final responses are far from sufficient for reliable measurement, and the measurement error in the final response-based latent trait estimates can propagate to the process-incorporated scores through the conditional expectation. Test developers are advised to employ a sufficiently large scoring set so that reliable response-based latent trait estimates can be obtained.

Second, the proposed process-incorporated scoring approach aimed at improving the measurement precision, or reliability, of the assessment through MSE reduction. The validity of the scoring rule, however, is a separate critical issue that deserves further attention. Looking at the empirical interpretations of the process-incorporated scores, it appeared that individuals were scored higher based on processes when they gave up on the track to a correct response, demonstrated partially correct responses, or slipped on the final response due to careless mistakes. In these cases, increasing the individuals' latent trait estimates may be reasonable, because each of these patterns constitutes evidence of partial or full mastery of the required skills for completing the tasks. Meanwhile, individuals who reached correct responses but with less efficient problem-solving strategies received lower process-incorporated scores. Assigning different trait estimates based on the choice of test-taking strategies may be more controversial: On the one hand, usage of more efficient problem-solving strategies was found positively correlated with the final score on other tasks in the empirical example, providing validity evidence on the use of such information for proficiency assessment. From this perspective, problem-solving strategy information may be incorporated into ability estimation similar to other types of collateral information, such as response times (e.g., Bolsinova & Tijmstra, 2018; van der Linden, 2007) and other covariates in

latent regression (e.g., von Davier et al., 2006). On the other hand, test takers may be unaware that they are scored based on information aside from task completion, raising concerns about the face validity and broader implications of the scoring criterion. At the same time, it raises test design questions such as whether examinees should be instructed that their scores can be affected by the problem-solving process. Evaluation of measurement validity would involve not only a search for empirical evidence that support the use and interpretation of the scores but also the appraisal of social consequences of score uses (Messick, 1989). At the same time, scoring algorithms based on process data bring forth new validity and equity questions, such as whether different subgroups of test takers behave similarly and how to evaluate potential algorithmic bias (e.g., Bejar et al., 2016; Zumbo & Hubley, 2017). We leave the question of how to best assist experts with the validation of data-driven latent proficiency estimators to future research.

Third, similar to item response modeling, the process-incorporated scoring rule should also be generalizable to the *population*. Conceptually, many high-dimensional regression methods search for a set of parameters that minimizes the expected loss at the population level instead of the finite sample of training data. In the presence of noise in the process data, this mandates an adequate sample size as well as variable selection to prevent overfitting. We recommend users adopt similar simulation studies as in the current study, by manipulating the number and prevalence of possible actions based on the actual item, to approximate the sample size requirement for implementing the procedures. Measures should also be taken to prevent overfitting when one extracts the sufficient statistic by fitting the high-dimensional regression model involving process features, for example, by using variable selection in conjunction with cross-validation.

Lastly, to establish the theoretical results on improved measurement efficiency, several assumptions were made in the current framework. One should note that, if some of the assumptions are violated, the efficiency results may be discounted, and several sources of bias may be introduced to the estimator. One of such instances is when the final score on an item is affected by systematic, construct-irrelevant variance. The presence of construct-irrelevant variance in the process alone does not constitute an assumption violation: We conducted a small-scale simulation study where the response process on each item was generated to depend on both θ and an additional, construct-irrelevant trait, $\eta \perp \theta$. As long as the final responses were independent of η, the process-incorporated proficiency estimator remained uncontaminated by η. Furthermore, we do expect that the proposed estimator improves upon the response-based estimator even beyond the three assumptions. Although practical diagnostic procedures are not yet available for verifying these assumptions by the data, in practice, practitioners could also perform an analysis as demonstrated in the Empirical Example section, to get a sense of the potential improvement in reliability by adopting the process-incorporated score.

4.3. Future Extensions

The methods for data-driven score refinement based on problem-solving processes can be extended in several ways. To start, while the current study provides one approach to increasing measurement reliability with process information, other methods, such as latent regression with process as covariates and confirmatory models with both response and process indicators, may be developed. Unlike the ordinal final outcomes which are designed to assess the measured trait, the problem-solving process data is high-dimensional and can contain substantial construct-irrelevant variance. For these parametric models, effective methods for variable selection will be needed to parse out the signal (θ-related information) from the "noise". Another potential extension of the process-incorporated scoring method is to diagnostic assessments (e.g., Rupp et al., 2010), where, instead of measuring individuals on the continuous proficiency continuum, the goal is to classify individuals into latent classes based on their mastery status of discrete skills. Lastly, we excluded observations with omissions in the current analysis, because missingness due to omissions often

requires careful treatment in both item calibration and scoring (e.g., Frey et al., 2018; Rose et al., 2017; Ulitzsch et al., 2020) which is beyond the scope of the current study. Omissions can entail either good-faith attempts at a question or a lack of motivation, in which case, simply treating it as wrong introduces construct-irrelevant variance to final response scoring. At the same time, process data introduce new opportunities for disentangling different types of omissions. Future research can examine how process data can be incorporated into the scoring of tests in the presence of omissions.

Acknowledgments

This research was supported in part by NSF Grants SES-1826540, SES-2119938, DMS-2015417 and 1633360. The authors would like to thank Educational Testing Service for providing the data.

Appendix: Proofs of Theorem 1 and Theorem 2

To prove Theorem 1, we establish the following lemma.

Lemma 1. *Let X be a nonconstant random variable, and $f(\cdot)$ and $g(\cdot)$ be strictly increasing functions. Suppose that $f(X)$ and $g(X)$ have finite second moments. Then,* $\text{Cov}\left(f(X), g(X)\right) > 0$.

Proof of lemma 1. Let Y be an independent and identically distributed (i.i.d.) copy of X. It is easy to verify the following identity

$$\text{Cov}\left(f(X), g(X)\right) = \frac{1}{2} E\left[\left(f(X) - f(Y)\right)\left(g(X) - g(Y)\right)\right]. \tag{10}$$

Clearly, for any x and y, $(f(x) - f(y))(g(x) - g(y)) \geq 0$, and " $=$" holds if and only if $x = y$. Since $P(X \neq Y) > 0$, the right-hand side of equation (10) must be positive. □

Proof of Theorem 1. By Assumption A2 (local independence),

$$T_{\mathbf{X}_{-j}} = E\left[\hat{\theta}_{Y_j} | \mathbf{X}_{-j}\right] = E\left[E\left[\hat{\theta}_{Y_j} | \mathbf{X}_{-j}, \theta\right] | \mathbf{X}_{-j}\right] = E\left[E\left[\hat{\theta}_{Y_j} | \theta\right] | \mathbf{X}_{-j}\right] = E\left[m_j(\theta) | \mathbf{X}_{-j}\right].$$

Due to Assumption A3 (exponential family), the posterior distribution of θ given $\mathbf{X}_{-j}$ depends on $\mathbf{X}_{-j}$ only through the sufficient statistic $T_j(\mathbf{X}_{-j})$. In fact,

$$T_{\mathbf{X}_{-j}} = E\left[m_j(\theta) | \mathbf{X}_{-j}\right] = G_j(T_j(\mathbf{X}_{-j})),$$

where $G_j(t) = E\left[m_j(\theta)|T_j(\mathbf{X}_{-j}) = t\right]$. Furthermore, by making use of the exponential family form in Assumption A3 and the simple exchange of order of differentiation and integration, we can show that

$$G_j'(t) = \mathrm{Cov}\left[m_j(\theta), \eta_j(\theta)|T_j(\mathbf{X}_{-j}) = t\right].$$

Since both m_j and η_j are strictly monotone, Lemma 1 implies that $G_j'(t)$ is strictly positive or negative for all t and, therefore, G_j is strictly monotone. In other words, there is a one-to-one mapping between $T_{\mathbf{X}_j}$ and $T_j(\mathbf{X}_{-j})$. □

Proof of Theorem 2. From Theorem 1, we know that $T_{\mathbf{X}_{-j}}$ is a sufficient statistic of $\mathbf{X}_{-j}$ for each j. Since $\hat{\theta}_{\mathbf{Y}}$ is a function of $\mathbf{Y}$ and $\sigma(\mathbf{Y}_{-j}) \subseteq \sigma(\mathbf{X}_{-j})$, the conditional distribution $\hat{\theta}_{\mathbf{Y}}|T_{\mathbf{X}_{-j}}, Y_j$ is free of θ. Therefore, we have $E[\hat{\theta}_{\mathbf{Y}}|T_{\mathbf{X}_{-j}}, Y_j, \theta] = E[\hat{\theta}_{\mathbf{Y}}|T_{\mathbf{X}_{-j}}, Y_j] = \hat{\theta}_{\mathbf{X}_{-j}}$. It follows from the well-known Rao–Blackwell theorem (Casella & Berger, 2002) that $\hat{\theta}_{\mathbf{X}_{-j}}$ reduces the conditional variance and

$$E[(\hat{\theta}_{\mathbf{X}_{-j}} - \theta)^2|\theta] \leq E[(\hat{\theta}_{\mathbf{Y}} - \theta)^2|\theta]$$

holds for every j and θ.
By Cauchy–Schwarz inequality, we get

$$E[(\hat{\theta}_{\mathbf{X}} - \theta)^2|\theta] \leq E\left[\frac{1}{J}\sum_{j=1}^{J}(\hat{\theta}_{\mathbf{X}_{-j}} - \theta)^2|\theta] \leq E[(\hat{\theta}_{\mathbf{Y}} - \theta)^2|\theta\right].$$

References

AERA, APA, and NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association American Psychological Association.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V.2. *The Journal of Technology, Learning and Assessment, 4*(3). Retrieved from https://ejournals.bc.edu/index.php/jtla/article/view/1650

Bejar, I. I., Mislevy, R. J., & Zhang, M. (2016). Automated scoring with validity in mind. In A. A. Rupp & J. P. Leighton (Eds.), *The Wiley handbook of cognition and assessment* (pp. 226–246). https://doi.org/10.1002/9781118956588.ch10

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), Statistical theories of mental test scores (pp. 397–479). Addison-Wesley.

Blackwell, D. (1947). Conditional expectation and unbiased sequential estimation. *The Annals of Mathematical Statistics, 18*(1), 105–110.

Bolsinova, M., & Tijmstra, J. (2018). Improving precision of ability estimation: Getting more from response times. *British Journal of Mathematical and Statistical Psychology, 71*(1), 13–38.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (Vol. 2). Duxbury.

Clauser, B. E., Harik, P., & Clyman, S. G. (2000). The generalizability of scores for a performance assessment scored with a computer-automated scoring system. *Journal of Educational Measurement, 37*(3), 245–261.

Evanini, K., Heilman, M., Wang, X., & Blanchard, D. (2015). Automated scoring for the toefl junior® comprehensive writing and speaking test. *ETS Research Report Series, 2015*(1), 1–11.

Fife, J. H. (2013). Automated scoring of mathematics tasks in the common core era: Enhancements to m-rater in support of cbal™ mathematics and the common core assessments. *ETS research report series, 2013*(2), i–35.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). Automated essay scoring: Applications to educational technology. In B. Collis & R. Oliver (Eds.), *Proceedings of EdMedia + Innovate Learning 1999* (pp. 939–944). Association for the Advancement of Computing in Education (AACE).

Frey, A., Spoden, C., Goldhammer, F., & Wenzel, S. F. C. (2018). Response time-based treatment of omitted responses in computer-based testing. *Behaviormetrika, 45*(2), 505–526.

He, Q., Veldkamp, B. P., Glas, C. A., & de Vries, T. (2017). Automated assessment of patients' self-narratives for posttraumatic stress disorder screening using natural language processing and text mining. *Assessment, 24*(2), 157–172.

He, Q., Veldkamp, B. P., Glas, C. A., & Van Den Berg, S. M. (2019). Combining text mining of long constructed responses and item-based measures: A hybrid test design to screen for posttraumatic stress disorder (ptsd). *Frontiers in Psychology, 10*, 2358.

He, Q., & von Davier, M. (2016). Analyzing process data from problem-solving items with N-grams: Insights from a computer-based large-scale assessment. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development* (pp. 750–777). IGI Global. https://doi.org/10.4018/978-1-4666-9441-5.ch029

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika, 30*(1/2), 81–93.

Kim, J. K., & Nicewander, W. A. (1993). Ability estimation for conventional tests. *Psychometrika, 58*(4), 587–599.

LaMar, M. M. (2018). Markov decision process measurement model. *Psychometrika, 83*(1), 67–88.

Lehmann, E. L., & Romano, J. P. (2005). *Testing statistical hypotheses* (3rd ed.). Springer.

Liu, H., Liu, Y., & Li, M. (2018). Analysis of process data of pisa 2012 computer-based problem solving: Application of the modified multilevel mixture irt model. *Frontiers in Psychology, 9*, 1372.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5–11.

Muraki, E. (1992). A generalized partial credit model: Application of an em algorithm. *ETS Research Report Series, 1992*(1), i–30.

OECD. (2012). *Literacy, numeracy and problem solving in technology-rich environments: Framework for the oecd survey of adult skills*. OECD Publishing.

Page, E. B. (1966). The imminence of grading essays by computer. *The Phi Delta Kappan, 47*(5), 238–243.

Qiao, X., & Jiao, H. (2018). Data mining techniques in analyzing process data: A didactic. *Frontiers in Psychology, 9*, 2231.

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*. Danish Institute for Educational Research.

Rose, N., von Davier, M., & Nagengast, B. (2017). Modeling omitted and not-reached items in irt models. *Psychometrika, 82*(3), 795–819.

Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of IntelliMetric^{TM} essay scoring system. *The Journal of Technology, Learning and Assessment, 4*(4). Retrieved from https://ejournals.bc.edu/index.php/jtla/article/view/1651

Rupp, A. A. (2018). Designing, evaluating, and deploying automated scoring systems with validity in mind: Methodological design decisions. *Applied Measurement in Education, 31*(3), 191–214.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.

Schleicher, A. (2008). Piaac: A new strategy for assessing adult competencies. *International Review of Education, 54*(5–6), 627–650.

Tang, X., Wang, Z., Liu, J., & Ying, Z. (2021a). An exploratory analysis of the latent structure of process data via action sequence autoencoders. *British Journal of Mathematical and Statistical Psychology, 74*(1), 1–33.

Tang, X., Zhang, S., Wang, Z., Liu, J., & Ying, Z. (2021b). Procdata: An R package for process data analysis. *Psychometrika, 86*(4), 1058–1083.

Tang, X., Wang, Z., He, Q., Liu, J., & Ying, Z. (2020). Latent feature extraction for process data via multidimensional scaling. *Psychometrika, 85*(2), 378–397.

Tikhonov, A. N. & Arsenin, V. Y. (1977). *Solutions of ill-posed problems* (pp. 1–30). New York.

Ulitzsch, E., von Davier, M., & Pohl, S. (2020). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response. *British Journal of Mathematical and Statistical Psychology, 73*, 83–112.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*(3), 287.

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (2006). 32 the statistical procedures used in national assessment of educational progress: Recent developments and future directions. *Handbook of Satistics, 26*, 1039–1055.

Wainer, H., Dorans, N. J. , Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). *Computerized adaptive testing: A primer*. Routledge.

Xu, H., Fang, G., Chen, Y., Liu, J., & Ying, Z. (2018). Latent class analysis of recurrent events in problem-solving items. *Applied Psychological Measurement*, 0146621617748325.

Zumbo, B. D., & Hubley, A. M. (2017). *Understanding and investigating response processes in validation research* (Vol 26). Springer.

Manuscript Received: 29 JUL 2020
Final Version Received: 12 JUL 2022