ACCURATE ASSESSMENT VIA PROCESS DATA

Susu Zhang, Zhi Wang, Jitong Qi, Jingchen Liu and Zhiliang Ying

columbia university

July 20, 2020

ACCURATE ASSESSMENT VIA PROCESS DATA

Abstract

Accurate assessment of 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 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 empirically that appropriately including such information in assessment will substantially improve relevant assessment precision. The precision is measured empirically by out-of-sample test reliability.

Key words: Process data; ability estimation; automated scoring; Rao-Blackwellization

1. Introduction

The main task of educational assessment is to provide reliable and valid estimates of students' abilities based on their responses to test items. Much of the effort in the past decades focused on the item response theory (IRT) models, the responses of which are often dichotomous (correct/incorrect), polytomous (partial score), and generically discrete (e.g. multiple choice item). The rapid advancement of information technology has enabled the collection of various sorts of process data for 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 that document the processes of test-takers' solving a problem contains 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, etc. (e.g., He & von Davier, 2016; LaMar, 2018; Liu, Liu, & Li, 2018; Xu, Fang, Chen, Liu, & Ying, 2018).

The emergence of process data provides the psychometric community with opportunities to develop cutting-edge research and, at the same time, brings forward challenges. Process data indeed contain rich information about the students. Much of the literature focuses on developing new research directions. In this paper, we take a different angle and try to answer the question how existing research could benefit from the analysis. In particular, we develop a method to incorporate information in process data to the scoring formula. There are two key features to consider: reliability and validity. For reliability, we show that the process-data-based assessment is significantly more accurate

than that based on the IRT models. In particular, we demonstrate through a real data analysis that the process-data-based scoring rule yields much higher reliability than that of the IRT-model-based ability estimates. Score based on a single process data item could be as accurate as that of three IRT-based scores. Furthermore, we also provide a theoretical framework under which process-data-based scores are guaranteed to yield more accurate estimates of students' abilities, of course, under certain conditions. Reliability, on the other hand, is a more complex problem. Process data record the entire problem-solving process that reveals different aspects of a student. It is unclear which part of the process is related to the particular ability of interest. It is conventionally up to the domain experts to identify construct-relevant process features and derive scoring rules. Such an approach is costly and cannot be scaled up. Our approach considers an automated scoring system of process data. Features are extracted through an exploratory analysis that typically lacks interpretation. We take advantage of the IRT scores that helps us to guide scoring rules to yield a valid test score. The entire procedure does not require particular knowledge of the item design.

Process data are often in a format not easy to directly incorporate to analysis. We preprocess the data by embedding each response process to a finite dimensional vector space. There are multiple methods to fill this task, including $n-$gram language modelling (He & von Davier, 2016), sequence-to-sequence autoencoders (Tang, Wang, Liu, & Ying, 2019), and multidimensional scaling (Tang, Wang, He, Liu, & Ying, 2019). In this paper, we use feature extracted from multidimensional scaling.

To the authors' best knowledge, this is the first piece of work to use process data to improve measurement accuracy. A literature that is remotely related to the currently work is the automated scoring systems for constructed responses, for example, 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, Laham, & Landauer, 1999; Page, 1966). In the context of computer-based problem solving items, process

features have been used to predict the final item response (e.g., Qiao & Jiao, 2018; Tang, Wang, He, et al., 2019; Tang, Wang, Liu, & Ying, 2019). Many automated scoring algorithms were shown to produce comparable scores to expert ratings.

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 two-step conditional expectation 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 processes.

The rest of the paper is organized as follows. Section 2 describes the statistical formulation and the two-step conditional expectation method for score refinement is introduced. Theoretical results on mean squared error (MSE) reduction in latent trait estimation, as well as illustrative example for practical use, are presented. Section 3 contains 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 that compares the proposed method to the original response-based scoring in several aspects. A discussion of the implications and limitations is provided in Section 4.

2. Latent Trait Estimation with Processes and Responses

We start with a generic framework for the proposed approach, followed by a more specific illustrative example. Section 2.1 lays the statistical foundation upon which the proposed approach is built. Section 2.2 describes the proposed two-step conditional expectation approach for process-based latent trait measurement and presents some related theoretical results. Section 2.3 contains a detailed illustrative example, which shows how the generic approach is implemented in practice.

2.1. Statistical Formulation

Consider a test of J items that is 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, 2012),

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

where $\boldsymbol{\zeta}_j$ is the parameter vector associated with 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), where

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

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

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

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 estimators 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 by $\mathbf{X}_j = (X_{j1}, \ldots, X_{jk}) \in \mathbb{R}^K$. 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 $\sigma-$algebra generated by the random variable. Intuitively, this requires the extracted features to preserve full information about the final score so that they can perfectly predict them. Since final outcomes are deterministically derived from response processes, they can always be added into extracted features to guarantee $\sigma(Y_j) \subseteq \sigma(\mathbf{X}_j)$. Feature extraction methods such as n-gram language modelling (e.g., He & von Davier, 2016; Qiao & Jiao, 2018), multidimensional scaling (MDS; Tang, Wang, He, et al., 2019), and recurrent neural network-based sequence-to-sequence autoencoders (Tang, Wang, Liu, & Ying, 2019), which have documented performance in terms of near-perfect final response prediction, can be applied in practice.

2.2. Two-Step Conditional Expectation Procedure

For a subset of items, $\mathcal{B} \subseteq \{1, \ldots, J\}$, denote by $\mathbf{X}_\mathcal{B}$ and $\mathbf{Y}_\mathcal{B}$, the examinee's vectors of features extracted from their action sequences and of their final responses, respectively. Let $\hat{\theta}_{Y_\mathcal{B}}$ be an estimator of latent trait θ based on $\mathbf{Y}_\mathcal{B}$ using, e.g., Equation (1) or (2). This

subsection proposes a new estimator of θ by developing a two-step conditional expectation procedure for score refinement. The new estimator makes use of the information from the action sequences and is shown to improve upon $\mathbf{Y}_{\mathcal{B}}$ in terms of reducing the mean squared error.

To construct the new estimator, consider two disjoint subsets of items, $\mathcal{B}_1$ and $\mathcal{B}_2$. From the above definitions, we have $\mathbf{X}_{\mathcal{B}_i}$, $\mathbf{Y}_{\mathcal{B}_i}$ and $\hat{\theta}_{Y_{\mathcal{B}_i}}$, $i = 1, 2$. Below describes our construction of a new estimator of θ, $\hat{\theta}_{X_{\mathcal{B}_1}}$.

Procedure 1 (Construction of New Estimator)**.** *Given final score-based estimators* $\hat{\theta}_{Y_{\mathcal{B}_1}}$, $\hat{\theta}_{Y_{\mathcal{B}_2}}$ *and process features* $\mathbf{X}_{\mathcal{B}_1}$*, we construct a new estimator* $\hat{\theta}_{X_{\mathcal{B}_1}}$ *through the following two conditional expectations.*

Step 1: Regress $\hat{\theta}_{Y_{\mathcal{B}_2}}$ *on* $\mathbf{X}_{B_1}$ *to obtain* $T_X = E[\hat{\theta}_{Y_{\mathcal{B}_2}}|\mathbf{X}_{\mathcal{B}_1}]$.

Step 2: Regress $\hat{\theta}_{Y_{\mathcal{B}_1}}$ *on* T_X *to obtain* $\hat{\theta}_{X_{\mathcal{B}_1}} = E[\hat{\theta}_{Y_{\mathcal{B}_1}}|T_X]$.

The resulting estimator, $\hat{\theta}_{X_{\mathcal{B}_1}}$, is the new estimator for latent trait θ based on both the responses and the processes on items in set $\mathcal{B}_1$. Note that step 1 regresses the latent trait estimate from final scores on $\mathcal{B}_2$ against the process features on $\mathcal{B}_1$ to obtain T_X, while step 2 regresses the latent trait estimate based on $\mathcal{B}_1$ final scores against T_X. Switching the roles of $\mathcal{B}_1$ and $\mathcal{B}_2$, we can similarly obtain $\hat{\theta}_{X_{\mathcal{B}_2}}$.

The proposed procedure improves estimation of the latent trait under some assumptions, which are to be presented. The first assumption requires the conditional expectation of $\hat{\theta}_{\mathcal{B}_2}$ given θ to be monotone increasing in θ. This assumption is satisfied by virtually all reasonable trait estimators.

A1: $m(\theta) = E\left[\hat{\theta}_{Y_{\mathcal{B}_2}} \mid \theta\right]$ is monotone in θ and has a finite second moment.

Next we assume that the examinees' responses to items in $\mathcal{B}_2$ are correlated with behavioral patterns on $\mathcal{B}_1$ only through the measured trait θ, not through other latent or

observed traits. Since process features can include rich information about respondents other than the measured trait, this assumption requires $Y_{\mathcal{B}_2}$ to be "good" in the sense that no differential item functioning (DIF) occurs. For example, the process features $\mathbf{X}_{\mathcal{B}_1}$ may well predict an examinee's age, but responses $Y_{\mathcal{B}_2}$ shall not differentiate young or old people as long as they have the same level of θ. However, we do allow $Y_{\mathcal{B}_2}$ to be very "rough" measurements, in other words, $\hat{\theta}_{Y_{\mathcal{B}_2}}$ might be biased and have large variance, as long as the monotonicity assumption A1 is required.

A2: Given latent trait θ, $\mathbf{Y}_{\mathcal{B}_2}$ and $\mathbf{X}_{\mathcal{B}_1}$ are independent.

Finally, for technical development, an exponential family assumption is imposed on process features. The natural parameter $\eta(\theta)$ is assumed to be monotone so that there is no identifiability issue for θ.

A3: The probability density function for features $\mathbf{X}_{\mathcal{B}_1}$ takes the following form

$$f(\mathbf{X}_{\mathcal{B}_1}|\theta) = \exp\left\{\eta\left(\theta\right) T(\mathbf{X}_{\mathcal{B}_1}) - A(\theta)\right\} h(\mathbf{X}_{\mathcal{B}_1}), \tag{3}$$

where $T(\mathbf{X}_{\mathcal{B}_1})$ is a sufficient statistic for θ and the natural parameter $\eta\left(\theta\right)$ is monotone in θ with a finite second moment.

Theorem 1 shows that the first step of our proposed procedure can summarize extracted features into sufficient statistics.

Theorem 1. *Under Assumptions A1-A3, T_X is a sufficient statistic of $\mathbf{X}_{\mathcal{B}_1}$ for θ.*

The proof of Theorem 1 is provided in the Appendix. From the sufficiency of T_X with respect to θ, it could be shown that step 2 reduces the MSE of $\hat{\theta}_{Y_{\mathcal{B}_1}}$ for estimating θ by taking conditional expectation with respect to this sufficient statistic. This result follows directly from the Rao-Blackwell theorem (Blackwell, 1947; Casella & Berger, 2002), and is stated in Theorem 2.

Theorem 2. *Under assumptions A1-A3, we have*

$$E[(\hat{\theta}_{X_{\mathcal{B}_1}} - \theta)^2|\theta] \leq E[(\hat{\theta}_{Y_{\mathcal{B}_1}} - \theta)^2|\theta] \quad \textit{for every } \theta. \tag{4}$$

Remark 1. *It follows directly from Theorem 2 that the MSE of $\hat{\theta}_{X_{\mathcal{B}_1}}$ for estimating θ is less than or equal to that of $\hat{\theta}_{Y_{\mathcal{B}_1}}$, uniformly across all examinees. This holds even when $\hat{\theta}_{Y_{\mathcal{B}_2}}$ has large bias and variance. If $\hat{\theta}_{Y_{\mathcal{B}_2}}$ has some known nice properties such as unbiasedness, then step 2 in Procedure 1 is optional. In that case, $T_X = E[\theta|\mathbf{X}_{\mathcal{B}_1}]$ is the posterior mean of θ and has the smallest possible MSE.*

Remark 2. *In practice, the explicit distributions of $\hat{\theta}_{Y_{\mathcal{B}_2}} \mid \mathbf{X}_{\mathcal{B}_1}$ and $\hat{\theta}_{Y_{\mathcal{B}_1}} \mid T_X$ are unknown, and thus the two conditional expectations, $E[\hat{\theta}_{Y_{\mathcal{B}_2}} \mid \mathbf{X}_{\mathcal{B}_1}]$ and $E[\hat{\theta}_{Y_{\mathcal{B}_1}} \mid T_X]$ in Procedure 1 can be approximated on finite samples using pretest data, for example, using multiple regression or regularized multiple regression models. Alternatively, deep neural networks can be fitted to approximate the nonlinear relationships between $\hat{\theta}_{Y_{\mathcal{B}_2}}$ and $\mathbf{X}_{\mathcal{B}_1}$, and similarly between $\hat{\theta}_{Y_{\mathcal{B}_1}}$ and T_X.*

Putting Theorem 2 in the measurement context, the MSE reduction for θ estimation intuitively translates to the reduction of the expected measurement error. In practice, the proposed approach can be applied to derive new scoring rules based on the pretest data. This process-based new scoring rule can replace the original response-based scoring rule to produce more reliable proficiency estimates in subsequent operational testing. It is also possible to administer only a subset of pretest items in operational testing under the refined scoring rule, which can potentially achieve comparable measurement precision to original final response-based scoring, but with fewer items.

2.3. Illustration

We now illustrate the proposed method through a specific setting. Consider a test of J items, administered to N (pretest) examinees. For examinee i on item j, let $\mathbf{S}_{ij}$ and Y_{ij} denote item-level process data and polytomous final responses, respectively. Suppose that the polytomous final responses are locally independent given the unidimensional latent trait θ and follow the Graded Response Model (GRM; Samejima, 2016)

$$1 - P(Y_{ij} < c \mid \theta_i) = P(Y_{ij} \geq c \mid \theta_i) = \frac{1}{1 + \exp[-(a_j\theta_i + d_{jc})]}, \quad c = 1, \ldots, C_j, \tag{5}$$

where the jth item has levels $0, 1, \ldots, C_j$ with parameters a_j and $\mathbf{d}_j = (d_{j1}, \ldots, d_{jC_j})$.

The following steps provide a roadmap to implement Procedure 1 on the pretest data to produce a new process-based scoring rule for estimating θ on the subset of items $\mathcal{B}_1$. Specifically, let $\mathcal{B}_2 = \bar{\mathcal{B}}_1 = \{1, \ldots, J\} \setminus \mathcal{B}_1$.

(1) IRT parameter estimation: On the final responses, $\mathbf{Y} = \{Y_{ij} : i = 1, \ldots, N, j = 1, \ldots, J\}$, fit the GRM to obtain the item parameters for each item j, $(\hat{a}_j, \hat{\mathbf{d}}_j)$, for example using marginal MLE (Bock & Aitkin, 1981).

(2) Process feature extraction: For item $j = 1, \ldots, J$, extract K-dimensional process features $\mathbf{X}_{1j}, \ldots, \mathbf{X}_{Nj}$ from the problem-solving processes $\mathbf{S}_{1j}, \ldots, \mathbf{S}_{Nj}$ for each test-taker. For instance, if MDS (Tang, Wang, He, et al., 2019) is applied for process feature extraction, this step obtains $\mathbf{X}_{1j}, \ldots, \mathbf{X}_{Nj} \in \mathbb{R}^K$ that minimizes

$$\sum_{i<i'} (d_{ii'} - \|\mathbf{X}_{ij} - \mathbf{X}_{i'j}\|)^2, \tag{6}$$

where $\|\cdot\|$ is the Euclidean distance, $d_{ii'} = d(\mathbf{S}_{ij}, \mathbf{S}_{i'j})$ is the dissimilarity between action sequences of test-takers i and i', $\mathbf{S}_{ij}$ and $\mathbf{S}_{i'j}$, based on the dissimilarity metric $d(\cdot)$, for example, the order-based sequence similarity metric (OSS; Gómez-Alonso &

Valls, 2008). Details on MDS feature extraction from process data can be found in Tang, Wang, He, et al. (2019).

(3) Set $\mathcal{B}_2$ trait estimation: For each examinee i, using the estimated item parameters in step (1), $\hat{\boldsymbol{\zeta}}_{\mathcal{B}_2} = \{(\hat{a}_j, \hat{\mathbf{d}}_j) : j \in \mathcal{B}_2\}$, estimate latent ability based on responses from all items in set $\mathcal{B}_2$, $\mathbf{Y}_{i,\mathcal{B}_2}$, for example using EAP in Equation (2). Denote this latent trait estimate for examinee i by $\hat{\theta}_{i,Y_{\mathcal{B}_2}}$.

(4) First conditional expectation: With the process features on items in $\mathcal{B}_1$, $\mathbf{X}_{1\mathcal{B}_1}, \ldots, \mathbf{X}_{N\mathcal{B}_1}$, where $\mathbf{X}_{i\mathcal{B}_1}$ is obtained by column-binding the features extracted from each item in set $\mathcal{B}_1$ in step (2), and the latent trait estimates from set $\mathcal{B}_2$, $\hat{\theta}_{1,Y_{\mathcal{B}_2}}, \ldots, \hat{\theta}_{N,Y_{\mathcal{B}_2}}$, fit a regression model for $\hat{\theta}_{Y_{\mathcal{B}_2}} \sim \mathbf{X}_{\mathcal{B}_1}$ to approximate $E[\hat{\theta}_{Y_{\mathcal{B}_2}} \mid \mathbf{X}_{\mathcal{B}_1}]$, for example, using ridge regression (Tikhonov & Arsenin, 1977). Denote the estimated regression function by f_1, then for $i = 1, \ldots, N$, obtain $\hat{E}[\hat{\theta}_{i,Y_{\mathcal{B}_2}} \mid \mathbf{X}_{i\mathcal{B}_1}] = f_1(\mathbf{X}_{i\mathcal{B}_1}) = T_{iX}$.

(5) Set $\mathcal{B}_1$ trait estimation: For each examinee i, using the estimated item parameters in step (1) and the responses on items in $\mathcal{B}_1$, $\mathbf{Y}_{i,\mathcal{B}_1}$, estimate latent ability (e.g., using EAP). Denote this estimate by $\hat{\theta}_{i,Y_{\mathcal{B}_1}}$.

(6) Second conditional expectation: With the output from the first conditional expectation (step (4)), $T_{1X}, \ldots, T_{NX}$, and the latent trait estimates from set $\mathcal{B}_1$ (step (5)), $\hat{\theta}_{1,Y_{\mathcal{B}_1}}, \ldots, \hat{\theta}_{N,Y_{\mathcal{B}_1}}$, fit a regression model for $\hat{\theta}_{Y_{\mathcal{B}_1}} \sim T_X$ to approximate $E[\hat{\theta}_{Y_{\mathcal{B}_1}} \mid T_X]$, for example, using simple linear regression. Denote the estimated regression function by f_2, then for $i = 1, \ldots, N$, $\hat{E}(\hat{\theta}_{i,Y_{\mathcal{B}_1}} \mid T_{iX}) = f_2(T_{iX}) = \hat{\theta}_{i,X_{\mathcal{B}_1}}$.

(7) $\hat{\theta}_{i,X_{\mathcal{B}_1}}$ is the estimate of θ based on the process data from items in $\mathcal{B}_1$.

Figure 1 illustrates the steps taken to construct the process-based latent trait estimator on set $\mathcal{B}_1$, $\hat{\theta}_{X_{\mathcal{B}_1}}$.

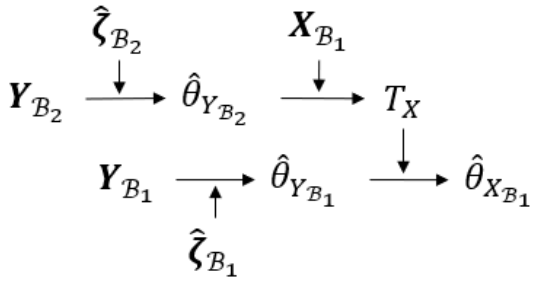

Figure 1: Flowchart for the construction of process-based latent trait estimator $\hat{\theta}_{X_{\mathcal{B}_1}}$.

At the operational testing stage, for each new subject $i^* \notin \{1, \ldots, N\}$, the following steps can be applied to obtain a latent trait estimate based on his or her action sequences on items in set $\mathcal{B}_1$:

(1) For each item j in $\mathcal{B}_1$, extract process features $\mathbf{X}_{i^*j}$ from action sequence $\mathbf{S}_{i^*j}$ using the same method as in step (2) above. When MDS is used for feature extraction, this translates to finding

$$\mathbf{X}_{i^*j} = \underset{\mathbf{X} \in \mathbb{R}^K}{\operatorname{argmin}} \sum_{i=1}^{N} (d_{ii^*} - \|\mathbf{X}_{ij} - \mathbf{X}\|)^2. \tag{7}$$

Column bind the extracted features $\mathbf{X}_{i^*j}$'s across items to obtain $\mathbf{X}_{i^*\mathcal{B}_1}$.

(2) Let $f_2 \circ f_1$ be the composition of the two conditional expectation functions obtained in steps (4) and (6), that is $f_2 \circ f_1(X) = f_2(f_2(X))$. Then the process-based latent trait estimate of examinee i^* on items in $\mathcal{B}_1$ is given by

$$\hat{\theta}_{i^*, X_{\mathcal{B}_1}} = f_2 \circ f_1(\mathbf{X}_{i^*\mathcal{B}_1}).$$

Remark 3. *Under the proposed scoring rule, $\mathcal{B}_1 \subsetneqq \{1, \ldots, J\}$ is a subset of the item pool, and thus the produced latent trait estimate only exploits the process information of a subset*

of available items. Consider a partition of the item pool into M disjoint subsets, $\{1, \ldots, J\} = \dot{\bigcup}_{m=1}^{M} \mathcal{B}'_m$, it is possible to implement the proposed scoring rule M times, each time with $\mathcal{B}_1 = \mathcal{B}'_m$ and $\mathcal{B}_2 = \bar{\mathcal{B}}'_m = \{1, \ldots, J\} \setminus \mathcal{B}'_m$, and obtain M estimates of θ, i.e., $\{\hat{\theta}_{X_{\mathcal{B}'_m}} : m = 1, \ldots, M\}$. Each $\hat{\theta}_{X_{\mathcal{B}'_m}}$ is the proficiency estimate based on the problem-solving processes on items in subset $\mathcal{B}'_m$. This way, the problem-solving processes of each item can be used. The production of an overall proficiency estimate for a test-taker based on multiple θ estimates, each from a different subset of tasks, is known as evidence accumulation under the evidence centered design framework (Mislevy, Almond, & Lukas, 2003). One simple way is to take the weighted sum of the individual estimators that minimizes the least squares deviation from the original response-based latent trait estimator. We leave the problem of best practices for evidence accumulation across tasks to future studies.

3. Empirical Example: PIAAC PSTRE

The proposed approach for score refinement is evaluated on the data collected from the problem-solving in technology-rich environments (PSTRE) assessment from the 2012 Programme for International Assessment of Adult Competency (PIAAC) survey. The empirical analyses are guided by two overarching objectives. First, the new process-based scoring rule is compared to the original final response-based scoring rule on the recovery of latent ability, so as to empirically validate the theoretical results on MSE reduction. Second, because process-based scoring and final response-based scoring are expected to produce different latent ability estimates of the same examinee, the current study further examines the problem-solving patterns associated with largest discrepancies in process- and response-based scores. In the following subsections, a description of the PIAAC PSTRE data is first provided, followed by the methods and findings from the empirical analyses.

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 age 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 PIAAC 2012 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 internet to obtain and evaluate information, communicate with others, and perform practical tasks (OECD, 2012). Successful completion of the PSTRE tasks thus 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 client, web browser, and spreadsheet. Test-takers were prompted to complete specific tasks in these interactive environments. Individuals' entire log of interactions with each item were recorded as log data. In addition, based on the extent of task completion, polytomous scores were derived for each item.

A sample item that resembles PSTRE tasks is shown in Figure 2. Test-takers 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 test-takers to identify, from the five web pages presented on the screen, all pages that do not require registration or fees and bookmark them. By clicking on each link, test-takers will be redirected to the corresponding website, where they can learn more about the website. For example, clicking

"Work Links" directs them to Figure 3, and further clicking on "Learn More" directs them to the page on Figure 4. 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 test-taker 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".

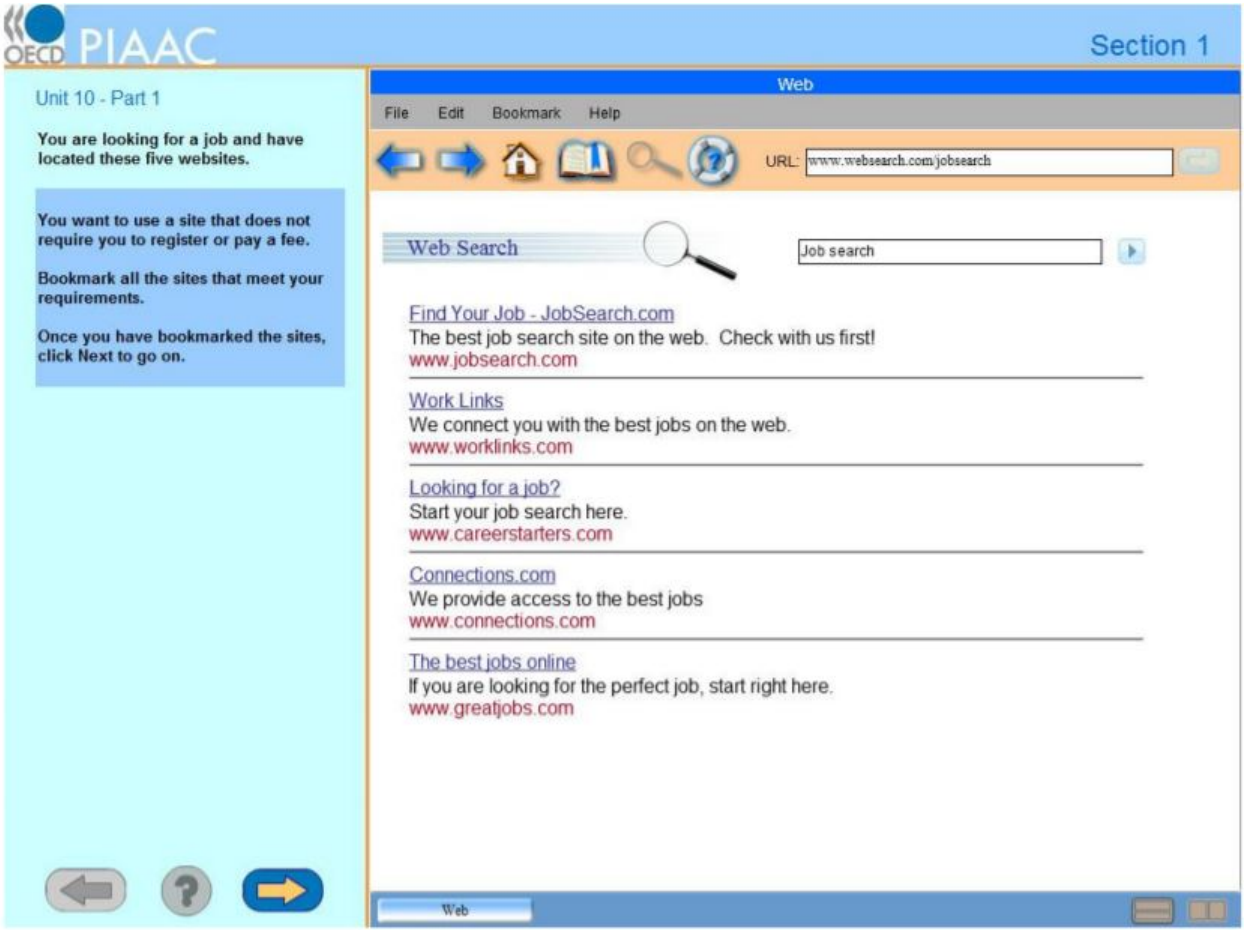

Figure 2: Home page of the PSTRE sample item.
Reprinted from OECD Sample Questions and Questionnaire.

The computer-based version of the 2012 PIAAC survey assigned each test-taker with two blocks of cognitive items, where each block consisted of fixed set items that assessed

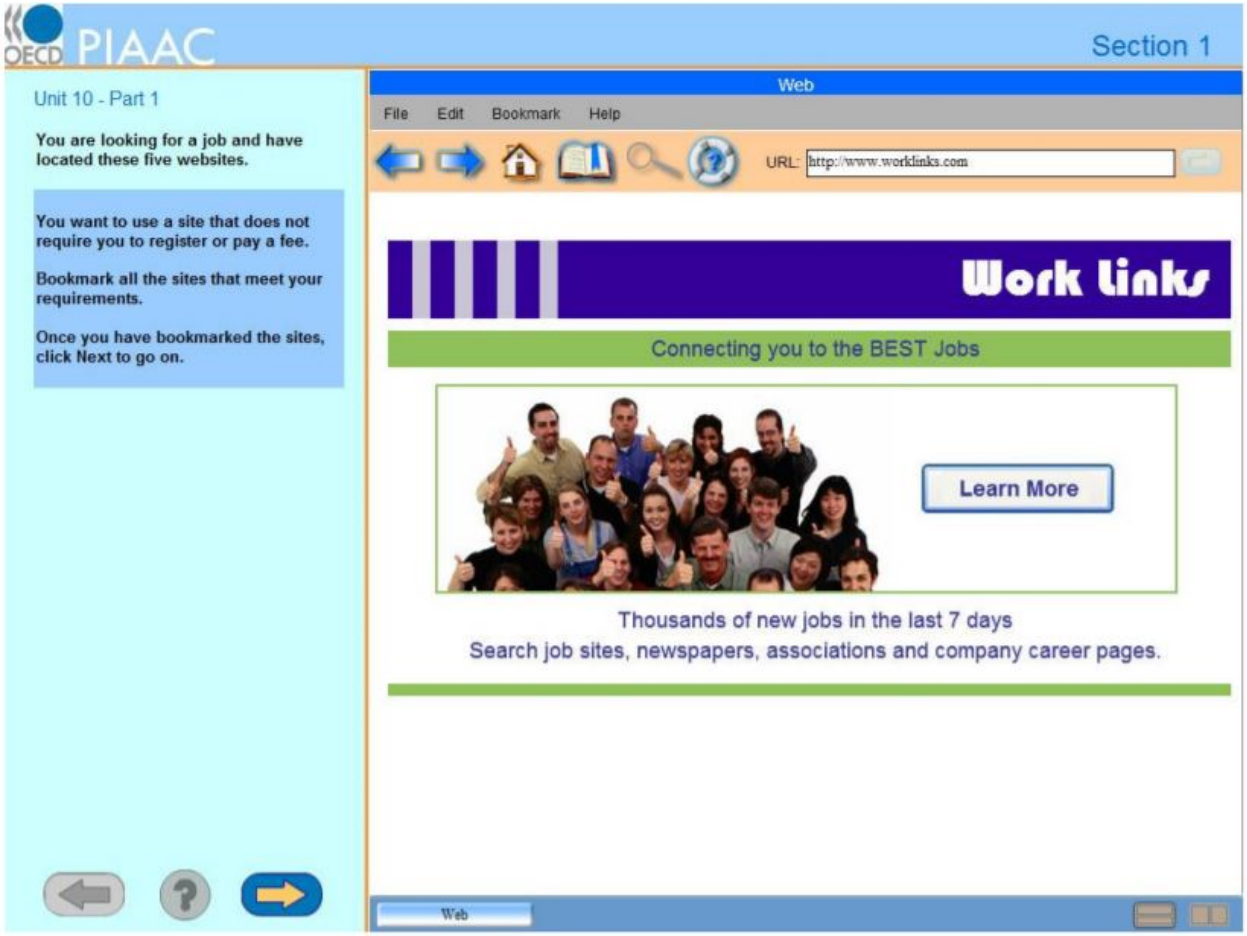

Figure 3: Web page returned from clicking the second link (i.e., "Work links") on the home page.

either literacy, numeracy, or PSTRE [1]. The current study used the PSTRE response and process data of individuals from five countries and regions, i.e., the United Kingdom (England and Northern Ireland), Ireland, Japan, the Netherlands, and the United States of America, and who were assigned to PSTRE for both blocks. 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 actions on the item and moved on to the next question [2]. This type of behavior can be regarded as

[1]Approximately one-sixth of test-takers were assigned to PSTRE block 1 (PS1) as the first block and PSTRE block 2 (PS2) as the second block.

[2]For the currently, we did not consider the time spent on the item by the test-takers and only look at the

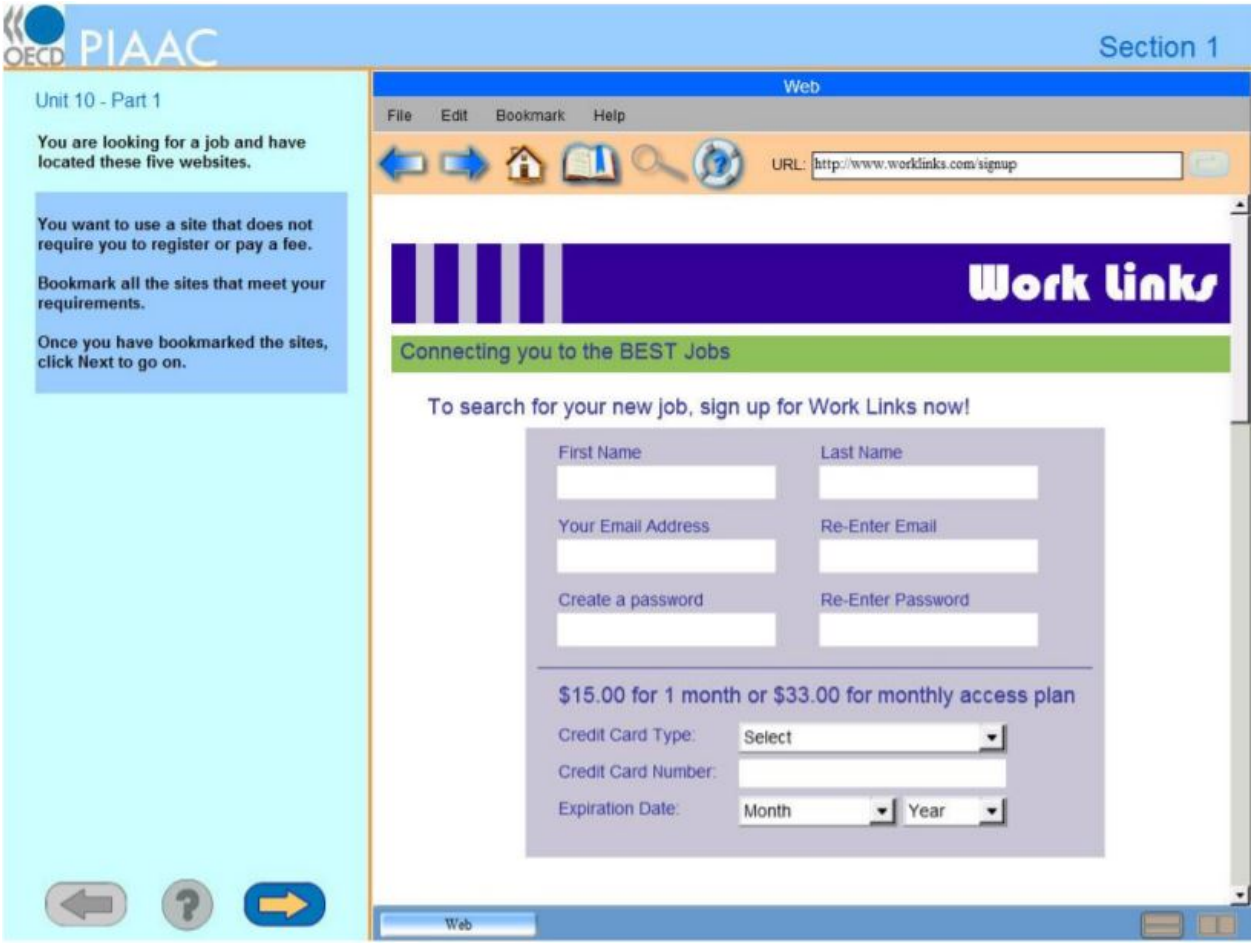

Figure 4: Web page returned from clicking "Learn More" on the "Work links" website.

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 sequences of each test-taker were recorded, and a polytomous final score calculated based on predefined scoring rubrics was available. These final scores (together with other demographic covariates) were used to estimate individuals' proficiency on PSTRE in the PIAAC survey. 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.

action sequences for differentiating individuals with or without omission.

Table 1: Descriptives information of the 14 PIAAC PSTRE items.

		Final Score			Sequence Length		
Item ID	Task name	Score levels	Median	Action types	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

Note. 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 performed by a subject.

3.2. Comparison of Process- and Score-based Estimators of Latent Proficiency

Ideally, one would want to compare the process- and response-based proficiency estimators in terms of their recovery of true latent ability. However, with empirical data, test-takers' true θs were unknown. The two proficiency estimators were instead compared on their agreement with performance on an external set of items that were designed to measure the same trait. Specifically, the 14 PSTRE items were randomly split into two sets of 7 items. The first set of 7 items, denoted by the scoring set $\mathcal{B}_s$, were used to implement response- or the process-based scoring rules and obtain the respective latent trait estimates $\hat{\theta}^{(s)}$. A separate latent trait estimates, $\hat{\theta}_Y^{(r)}$, were obtained from the polytomous responses on the second set of 7 items, denoted by the reference set $\mathcal{B}_r$. Any $\hat{\theta}^{(s)}$ obtained from the scoring set does not use reference set response information, and $\hat{\theta}_Y^{(r)}$ serves as an external criteria for evaluating different $\hat{\theta}^{(s)}$s. A particular $\hat{\theta}^{(s)}$ obtained from the scoring set was evaluated with two evaluation indices, namely the mean-squared deviation (MSE) from

$\hat{\theta}_Y^{(r)}$, i.e.,

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

and its Kendall's rank correlation (τ; Kendall, 1938) with $\hat{\theta}_Y^{(r)}$. Note that, unlike the true θ, $\hat{\theta}_Y^{(r)}$ was estimated based on final responses to only 7 items and was expected to contain significant measurement error. The correlation between $\hat{\theta}^{(s)}$ and $\hat{\theta}_Y^{(r)}$ is hence attenuated by the reliability of $\hat{\theta}_Y^{(r)}$, and the MSE of $\hat{\theta}^{(s)}$ with respect to $\hat{\theta}_Y^{(r)}$ 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 ($\hat{\theta}_Y^{(r)}$). Lower MSE and higher Kendall's τ hence indicate higher reliability.

Throughout the study, the EAP estimator (Equation (2)) was adopted for all response-based latent trait estimation, not only because it can handle all-correct or all-incorrect responses but also because the posterior mean (EAP) minimizes the posterior MSE. The prior distribution for θ was chosen to be the standard normal distribution, which was also used as the latent trait distribution in the marginal MLE item parameter estimation. Process features ($\mathbf{X}$s) were extracted from the action sequences using the `ProcData` R package (Tang, Zhang, Wang, Liu, & Ying, 2020). The OSS was chosen as the dissimilarity function, and the dimension of the MDS features for each item was set to $K = 30$.

The relationship between test length and agreement with $\hat{\theta}_Y^{(r)}$ was also examined. Using response-based scoring, test lengths of $t = 1$ to 7 items on the scoring set ($\mathcal{B}_s$) were considered. For a particular test length t, a response-based latent trait estimate ($\hat{\theta}_{Y_{\mathcal{B}_1}}$) was obtained on every possible combination ($\mathcal{B}_1$) of items of size t in $\mathcal{B}_s$. Because the proposed process-based scoring procedures require setting aside a set of items ($\mathcal{B}_2$) for the first

conditional expectation, only test lengths of $t = 1$ to 6 were considered for process-based scoring. For each t, every possible combination of t items was set as $\mathcal{B}_1$, the remaining $7 - t$ items in the scoring set ($\mathcal{B}_s \setminus \mathcal{B}_1$) were set as $\mathcal{B}_2$, and a process-based latent trait estimate ($\hat{\theta}_{X_{\mathcal{B}_1}}$) was obtained using the two-step conditional expectation procedures for each of the $\binom{7}{t}$ $\mathcal{B}_1$s. Linear models were adopted for both conditional expectations in the procedure. For the first conditional expectation, $E[\hat{\theta}_{Y_{\mathcal{B}_2}} \mid \mathbf{X}_{\mathcal{B}_1}]$, a ridge regression on MDS features was fitted to reduce overfitting using the `glmnet` (Friedman, Hastie, & Tibshirani, 2009) R package. The shrinkage parameter was tuned to minimize the 10-fold cross-validation error (deviance). The second conditional expectation, $E[\hat{\theta}_{Y_{\mathcal{B}_1}} \mid T_X]$, where T_X is the one-dimensional output of the first conditional expectation, was approximated using ordinary least squares regression.

Five-fold cross-validation was implemented to evaluate the behavior of different latent trait estimators on an independent sample, i.e., a group of new test-takers that were not used to develop the scoring rule. Specifically, the 2304 participants were randomly split into 5 equal-sized groups. In each fold, one group was set aside as the test set for evaluations, and the rest of the participants were used as the pretest sample to (1) estimate the GRM item parameters of the 14 items, (2) train the MDS for action sequences (Equation (6)), and (3) fit the two conditional expectation models for process-based scoring. On the test set, each test-taker's MDS process features were obtained based on their sequence dissimilarities with the pretest samples (Equation (7)), response-based latent trait estimates were obtained with item parameters calibrated from the pretest sample, and the two regression models obtained from pretest data were used to produce process-based latent trait estimates.

There are $\binom{14}{7}$ ways to partition 14 items into scoring and reference sets ($\mathcal{B}_s, \mathcal{B}_r$), each containing a different combination of items that vary in process and response informativeness. In the current study, 30 out of the $\binom{14}{7}$ partitions were randomly sampled

to obtain an approximate distribution of MSE and τ under each test length (up to 7 items) using response-based and process-based scoring. For instance, with process-based scoring using t items, the average MSE on a given partition, $(\mathcal{B}_s, \mathcal{B}_r)$, is

$$\bar{MSE}_t = \frac{1}{\binom{7}{t}} \sum_{\substack{\forall \mathcal{B}_1 \subset \mathcal{B}_s: \\ |\mathcal{B}_1| = t}} \left[\frac{1}{5} \sum_{f=1}^{5} MSE^f(\hat{\theta}_{X_{\mathcal{B}_1}}) \right],$$

where $MSE^f(\hat{\theta}_{X_{\mathcal{B}_1}})$ is the MSE evaluated on the fth fold in the cross-validation, given by

$$MSE^f(\hat{\theta}_{X_{\mathcal{B}_1}}) = \frac{1}{|\mathcal{N}_f|} \sum_{i \in \mathcal{N}_f} (\hat{\theta}_{i,X_{\mathcal{B}_1}} - \hat{\theta}^{(r)}_{i,Y})^2.$$

Here, $\mathcal{N}_f$ is the set of participants in the fth fold, $\hat{\theta}_{i,X_{\mathcal{B}_1}}$ is individual i's latent trait estimate based on the processes on $\mathcal{B}_1$, and $\hat{\theta}^{(r)}_{i,Y}$ is individual i's latent trait estimate based on the 7 polytomous responses on the reference set $\mathcal{B}_r$. Similar to the calculation of MSE, we can also evaluate by the average Kendall's τ for different test lengths. The MSE and τ for response-based estimators can be obtained in the same manner.

Figure 5 displays the box plots the average MSEs for each partition of scoring and reference sets, using different number of items (t) for scoring (x-axis, number of items in Set 1) and different estimators. The green boxes correspond to the response-based θ estimates with respect to reference set proficiency, and the red boxes correspond to that of the process-based θ estimates. Each box plot represents the distribution of the $\bar{MSE}$s of the 30 partitions for a particular test length and estimator. The horizontal dashed line gives the averaged MSE of the response-based latent estimate using all 7 items across all 30 partitions. One can observe that for all test lengths between 1 and 6 items, the process-based latent trait estimates consistently demonstrated smaller MSE, indicating higher agreement with the performance on an external set of similar tasks (i.e., the reference set response-based latent trait estimate). The advantage of the process-based

estimator was more salient for shorter tests and diminished as the test length increased. In particular, with one item ($t = 1$), the process-based $\hat{\theta}_{X_{\mathcal{B}_1}}$ achieved similar median MSE with final response-based $\hat{\theta}_{Y_{\mathcal{B}_1}}$ using 3 items ($t = 3$), and with $t = 2$, the median MSE of $\hat{\theta}_{X_{\mathcal{B}_1}}$ was comparable to that with 6 items using final responses. With 3 or more items, the process-based $\hat{\theta}_{X_{\mathcal{B}_1}}$s consistently achieved similar or lower MSE than the response-based estimators using all 7 items in the scoring set on all 30 partitions. Although the MSEs were calculated with respect to the estimated $\hat{\theta}$s on the reference sets, the lower MSEs using process-based scoring appeared to be consistent with the theoretical results on the MSE reduction using the proposed two-step conditional expectation approach.

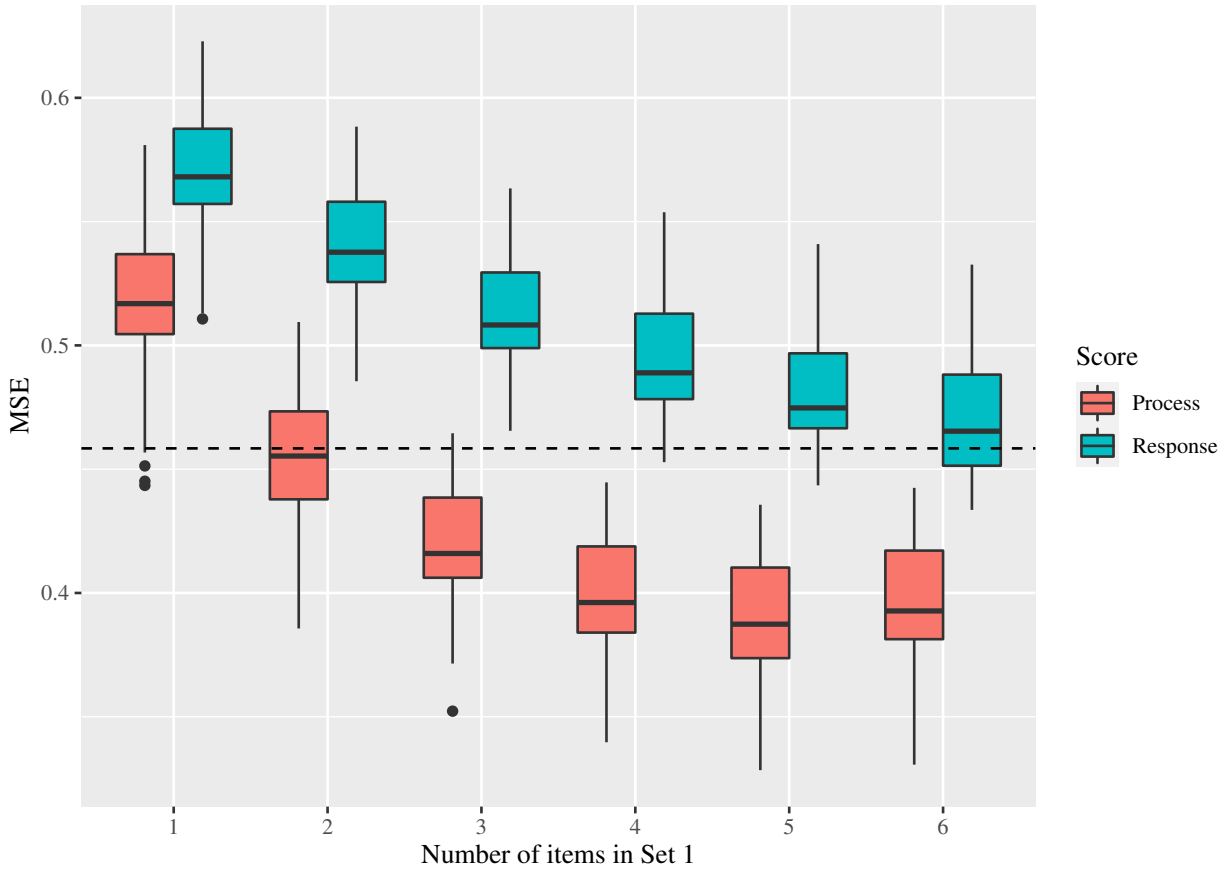

Figure 5: Distribution of mean-squared deviation (MSE) between reference set θ estimate and Set $\mathcal{B}_1$ θ estimate across different splits of scoring and reference sets, based on different number of Set $\mathcal{B}_1$ items.

The box plots for Kendall's rank correlations (τs) between reference set performance

and different estimates from the scoring set are presented in Figure 6. Unlike the MSE, which reflects absolute deviations, τ reflects the strength of associations of the two θ estimates in the relative ranking of test-takers. It could be observed that the correlations with reference set performance were consistently larger using process-based 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. In addition, scores based on processes required less items to achieve a given level of agreement. For instance, Kendall's τ based on $\hat{\theta}_{X_{\mathcal{B}_1}}$ with two items was similar to that of $\hat{\theta}_{Y_{\mathcal{B}_1}}$ with 4 items.

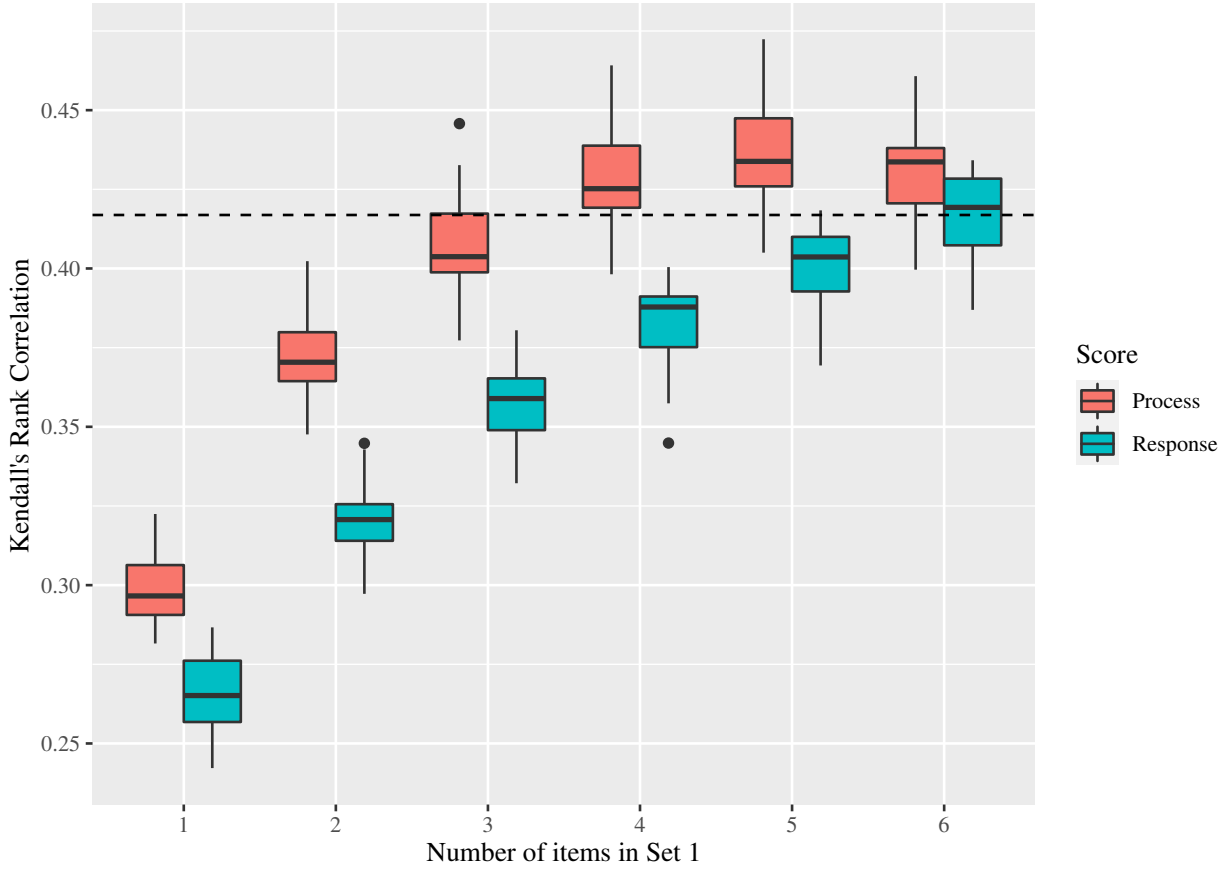

Figure 6: Distribution of Kendall's rank correlations between reference set θ estimate and Set $\mathcal{B}_1$ θ estimate across different splits of scoring and reference sets, based on different number of Set $\mathcal{B}_1$ items.

The above evaluations focused on the agreement between reference set performance

and process- vs. response-based θ estimates across all examinees. One may also be interested in how the two θ estimation 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. Using the same 30 partitions of scoring and references sets from above, we compared the $\hat{\theta}_{Y_{\mathcal{B}_1}}$s and $\hat{\theta}_{X_{\mathcal{B}_1}}$s produced using 6 items[3]. On a test set, for each pair of response- and process-based estimators, $\hat{\theta}_{Y_{\mathcal{B}_1}}$ and $\hat{\theta}_{X_{\mathcal{B}_1}}$, we regressed $\hat{\theta}_{Y_{\mathcal{B}_1}}$ on $\hat{\theta}_{X_{\mathcal{B}_1}}$ using ordinal least squares regression and calculated 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 to some extent reflect the relative difference in performance rankings based on $\hat{\theta}_{Y_{\mathcal{B}_1}}$ and $\hat{\theta}_{X_{\mathcal{B}_1}}$. The two ends of the deciles contain individuals whose process- and response-based latent trait estimates disagree the most in terms of rankings: For individuals in the first decile, their performance rankings based on responses were expected to be much lower than that based on their problem-solving processes. Individuals in the 10th decile were ranked higher based on final responses than based on processes. Individuals closer to the middle (4th - 6th decile) were expected to have relatively similar rankings based on processes and responses.

The box plots of the average MSEs with respect to reference set performance $\hat{\theta}_Y^{(r)}$ across the 30 partitions, separated by test-set individual's residual deciles, are shown in Figure 7. It can be observed that, when the two scores mostly agree on individuals' rankings, the MSEs of $\hat{\theta}_{Y_{\mathcal{B}_1}}$ and $\hat{\theta}_{X_{\mathcal{B}_1}}$ with respect to reference set performance were very similar. However, when the two scores disagreed the most (i.e., 1st, 2nd, 9th, and 10th deciles), the MSEs of process-based θ estimates were much lower than that of response-based estimates. Intuitively, the process- and response-based estimators can be

[3]There were again $\binom{7}{6}$ possibilities to select 6 items for $\mathcal{B}_1$ from the 7 scoring set items. Similar to the earlier evaluations, the reported performance results averaged across all possible combinations.

thought of as two judges, one judging individuals' performance based on the problem-solving processes, and the other judging solely based on the final outcome. The advantage of the process-based scores on the two extremes of residual deciles suggest that, when the two judges disagree the most, the process-based judge's estimate of an examinee's proficiency consistently predicted performance on other similar tasks better.

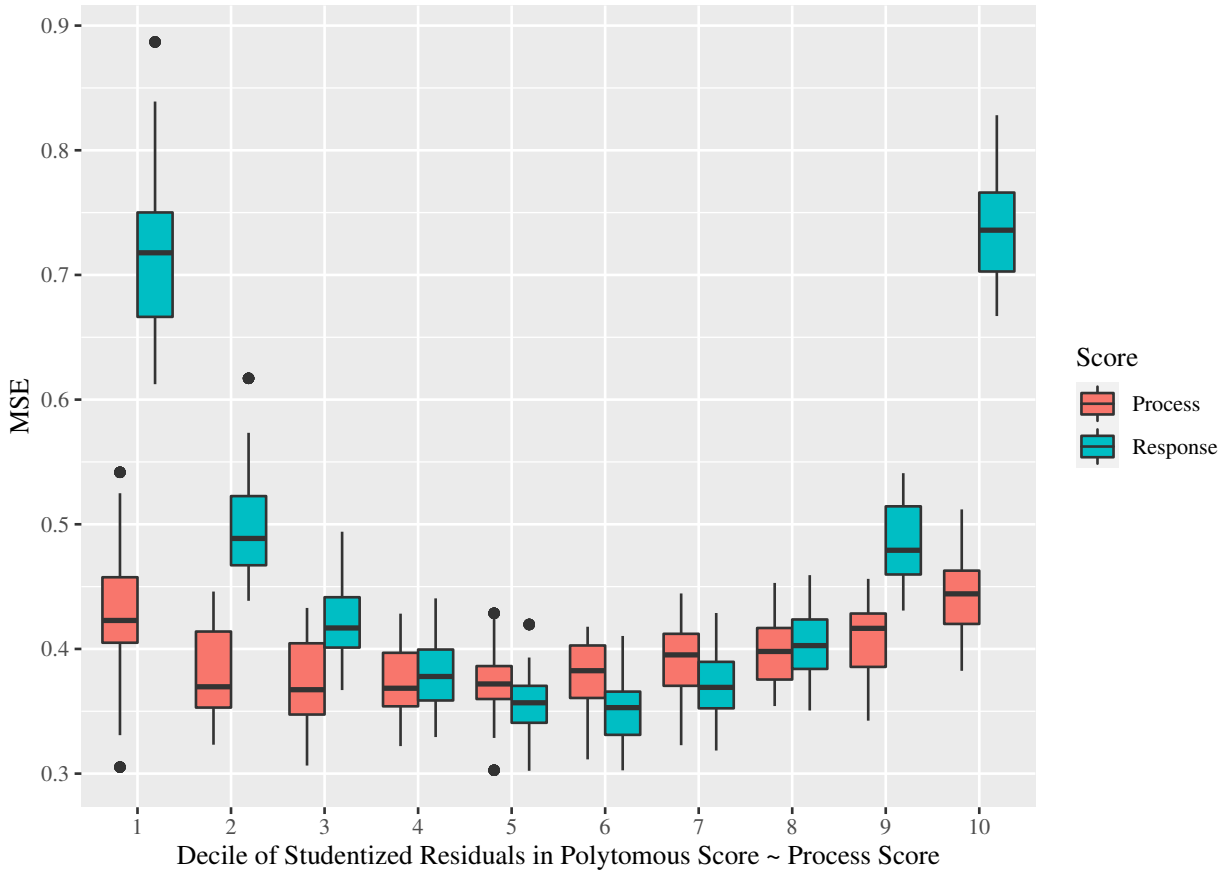

Figure 7: Distribution of MSE with reference set latent trait estimate in each residual decile.

3.3. Empirical Interpretations of Process-based Scores

The evaluation results suggested that the proposed process-based latent trait estimate procedures led to an increase in consistency with performance on an external set of items, and that the improvement appeared most significant for individuals whose process-based and response-based latent trait estimates disagreed most. One question worth asking is

how the process-based approach scores individuals differently from the response-based approach. We explored this question by looking at the sequences of individuals whose process- and response-based latent trait estimates disagree the most, i.e., have the highest or lowest Studentized residuals for $\hat{\theta}_{Y_{\mathcal{B}_1}} \sim \hat{\theta}_{X_{\mathcal{B}_1}}$.

Since the purpose of this subsection is interpretation rather than performance evaluation, we tried to include as many items and participants as possible. Only a single item U06a was chosen as $\mathcal{B}_2$, due to its shorter sequence lengths and lower sequence diversity. All the other 13 items constitute the set $\mathcal{B}_1$. In addition, all 2304 participants were included to obtain the process- or response-based latent trait estimates without setting aside a testing set. For the individuals in the bottom and top 10 in the Studentized residuals, we visually examined their action sequences on the 13 items.

Figure 8 shows the scatterplot of the latent trait estimates produced using processes (x-axis) and using responses (y-axis). Crosses and squares correspond to individuals with the lowest and highest Studentized residuals, respectively. The ten individuals with the lowest Studentized residuals, i.e., those who received higher ranking based on processes, were mostly at the bottom of the response-based proficiency continuum. In other words, their final responses to the questions were mostly incorrect. However, based on the problem-solving processes, they were placed in the middle/mid-low region of the proficiency continuum. Below are some patterns that were identified from these individuals' action sequences. (Note: Examinees 1 - 10 represent persons with lowest to 10th lowest Studentized residuals.)

- Give-up/partial response: The examinees performed some of the key steps for one or more questions but clicked "Next, Next_OK" before reaching a credited response. For example, on U16, which requires sending an email to a list of recipients containing some key information, examinee 1 copied down the key information, opened the "reply to all" window for sending emails, but decided to proceed to next question

without sending out the email. On the same question, examinees 5 and 6 both typed the key email content and recipients, but proceeded to the next question without sending the email.

- Partial mastery: The examinees performed most of the key steps for one or more questions but missed the credited response. For example, on item U04a which requires creating a spreadsheet, examinee 2 managed to make a spreadsheet with the key elements (column and row names, numerical entries) but was incorrect on some numerical entries. On U23 (making a request to return a lamp), examinee 2 managed to submit the return but selected the wrong reason for the return. On U16, examinee 4 sent out the correct email to one recipient but did not cc the rest. On item U21 that requires making ticket reservations in the browser, examinee 7 successfully reserved tickets, but the tickets do not meet the requirements by the question. On question U02 that requires making room reservations, examinees 9 and 10 made a few room reservations but with incorrect starting or ending times.

- Careless mistakes: The examinees demonstrated the required skills for completing the task but slipped due to careless mistakes. For example, on item U11b, which requires sorting emails in a particular folder, examinees 3 and 8 sorted the emails in the Inbox (default, wrong folder).

On the other side, the ten individuals with the highest Studentized residuals, i.e., those who received lower ranking based on processes, were mostly at the top of the response-based proficiency spectrum. Their proficiencies based problem-solving processes, however, were in the middle/mid-high range. For most of these individuals, there were questions that they successfully completed but with less efficient methods. For example, on item U16, individuals could send the email to the correct group recipients using "reply to all" and copy/paste the key information to the email contents. Several examinees typed

the recipients and the email content themselves instead. Another example is U19a, where examinees were required to identify one row in a long spreadsheet and extract relevant information. While they can directly locate the row using "search", some examinees visually inspected the entire spreadsheet to find the row. Aside from inefficient strategies, a few examinees also performed large number of redundant steps on some questions that were not required for successful task completion.

4. Discussions

Problem-solving processes contain rich information on individuals characteristics, including the measured construct. The current study introduces a method to refine final response-based latent trait estimates using the additional information from problem-solving processes. A two-step conditional expectation approach was proposed for the score refinement. Aside from choosing an appropriate IRT model for the final responses, the proposed approach is completely data-driven and does not require prior specification of a latent trait model for the problem-solving processes. Therefore, its implementation can be mostly automated and will not require excessive expert input.

The main theorem states that, under some regularity conditions, the proposed approach can lead to MSE reduction in latent trait estimation compared to the original response-based estimation. An empirical study using the PIAAC PSTRE data further showed that the process-based latent trait estimates tended to have higher agreement with performance on similar tasks, thus higher reliability, compared to the response-based estimates. 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.

The current study demonstrated that the proposed process-based score refinement approach could consistently improve test reliability (i.e., achieving lower MSE and higher

Kendall's τ with the other half). However, there are a few caveats for its implementations, especially for exams with higher stakes. First, the current study evaluated the performance of the process-based and response-based latent trait estimators 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-based scores through the conditional expectations. Test developers are advised to have a sufficiently large scoring set, so that relatively reliable response-based latent trait estimates can be obtained. Second, the proposed process-based scoring approach only 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 to be addressed. Looking at the empirical interpretations of the process-based 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 demonstrated partial or full mastery of the required skills for completing the tasks. On the other hand, individuals who reached correct responses but with less efficient problem-solving behavior received lower process-based proficiency estimate. Although the adjusted scores were closer to the performance on similar tasks (Figure 7), penalizing test-takers based on inefficient test-taking strategies may be more controversial. Evaluation of measurement validity would require both substantive knowledge and expert input. We leave the question of how to best assist experts with the validation of data-driven latent proficiency estimators to future research.

The methods for data-driven score refinement based on problem-solving processes can

be extended in several ways. Because the proposed process-based scoring approach requires setting aside one or more items for the first conditional expectation, there will be at least one item whose information is not used for process-based scoring. To fully exploit the information from all items, the process-based scoring approach can be applied iteratively to different subsets of items to produce multiple latent trait estimators. One key question to be addressed is how multiple latent trait estimators based on different sets of tasks can be efficiently combined to produce an overall latent ability estimate. Another potential extension of the process-based scoring method is to diagnostic assessments (e.g., Templin, Henson, 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.

The proposed approach for process-based scoring can be particularly useful for low-stakes computer-based assessment scenarios, when the administration of long tests is unrealistic or burdensome. In such cases, with the additional information from problem-solving processes, the tests can be significantly shortened without sacrificing measurement reliability. An example is interim formative assessments during the learning process, where, after every one or a few classes, the educators may want to learn how well the students have mastered the recently taught contents. Administration of a long test after each several classes can be very burdensome for the students and may interrupt the learning process. In such cases, 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, Dorans, Flaugher, Green, & Mislevy, 2000) can also reduce 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-based 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 two conditional expectation models. With a pretest sample size of approximately 1843 examiness (80% of 2304 samples), the empirical studies based on the PIAAC PSTRE data demonstrated improvement in reliability for new subjects' proficiency estimates.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*, *4*(3).

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

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, *46*(4), 443–459.

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

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). Automated essay scoring: Applications to educational technology. In *Edmedia+ innovate learning* (pp. 939–944).

Friedman, J., Hastie, T., & Tibshirani, R. (2009). glmnet: Lasso and elastic-net regularized generalized linear models. *R package version*, *1*(4).

Gómez-Alonso, C., & Valls, A. (2008). A similarity measure for sequences of categorical data based on the ordering of common elements. In *International conference on modeling decisions for artificial intelligence* (pp. 134–145).

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 *Handbook of research on technology tools for real-world skill development* (pp. 750–777). IGI Global.

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.

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*.

Lord, F. M. (2012). *Applications of item response theory to practical testing problems*. Routledge.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series*, *2003*(1), i–29.

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

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.

Samejima, F. (2016). Graded response models. In *Handbook of item response theory, volume one* (pp. 123–136). Chapman and Hall/CRC.

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

Tang, X., Wang, Z., He, Q., Liu, J., & Ying, Z. (2019). Latent feature extraction for process data via multidimensional scaling. *arXiv preprint arXiv:1904.09699*.

Tang, X., Wang, Z., Liu, J., & Ying, Z. (2019). An exploratory analysis of the latent structure of process data via action sequence autoencoder. *arXiv preprint arXiv:1908.06075*.

Tang, X., Zhang, S., Wang, Z., Liu, J., & Ying, Z. (2020). Procdata: An R package for process data analysis. *arXiv preprint arXiv:2006.05061*.

Templin, J., Henson, R. A., et al. (2010). *Diagnostic measurement: Theory, methods, and*

applications. Guilford Press.

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

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.

Appendix: Proofs of Theorems 1 and 2

To prove Theorem 1, we need 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* $\mathrm{Cov}\left(f(X), g(X)\right) > 0$.

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

$$\mathrm{Cov}\left(f(X), g(X)\right) = \frac{1}{2} E\left[\left(f(X) - f(Y)\right)\left(g(X) - g(Y)\right)\right]. \tag{9}$$

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 (9) must be positive. □

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

$$T_X = E\left[\hat{\theta}_{Y_{\mathcal{B}_2}} | \mathbf{X}_{\mathcal{B}_1}\right] = E\left[E\left[\hat{\theta}_{Y_{\mathcal{B}_2}} | \mathbf{X}_{\mathcal{B}_1}, \theta\right] | \mathbf{X}_{\mathcal{B}_1}\right] = E\left[E\left[\hat{\theta}_{Y_{\mathcal{B}_2}} | \theta\right] | \mathbf{X}_{\mathcal{B}_1}\right] = E\left[m(\theta) | \mathbf{X}_{\mathcal{B}_1}\right].$$

Due to Assumption A3 (exponential family), the posterior distribution of θ given $\mathbf{X}_{\mathcal{B}_1}$ depends on $\mathbf{X}_{\mathcal{B}_1}$ only through the sufficient statistic $T(\mathbf{X}_{\mathcal{B}_1})$. In fact,

$$T_X = E\left[m(\theta) | \mathbf{X}_{\mathcal{B}_1}\right] = G(T(\mathbf{X}_{\mathcal{B}_1})),$$

where $G(t) = E\left[m(\theta) | T(\mathbf{X}_{\mathcal{B}_1}) = t\right]$. 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

$$G'(t) = \mathrm{Cov}\left[m(\theta), \eta(\theta) | T(\mathbf{X}_{\mathcal{B}_1}) = t\right].$$

Since both m and η are strictly monotone, Lemma 1 implies that $G'(t)$ is strictly positive or negative for all t and, therefore, G is strictly monotone. In other words, there is a one-to-one mapping between T_X and $T(\mathbf{X}_{\mathcal{B}_1})$. □

Proof of Theorem 2. From Theorem 1, we know that T_X is a sufficient statistic. Since $\hat{\theta}_{Y_{\mathcal{B}_1}}$ is $\sigma(\mathbf{X}_{\mathcal{B}_1})$ measurable, we have $E\left[\hat{\theta}_{Y_{\mathcal{B}_1}}|T_X, \theta\right] = E\left[\hat{\theta}_{Y_{\mathcal{B}_1}}|T_X\right] = \hat{\theta}_{X_{\mathcal{B}_1}}$. Theorem 2 then follows from the well-known Rao-Blackwell Theorem (Casella & Berger, 2002), noting that $\hat{\theta}_{X_{\mathcal{B}_1}}$ and $\hat{\theta}_{Y_{\mathcal{B}_1}}$ have the same bias. □

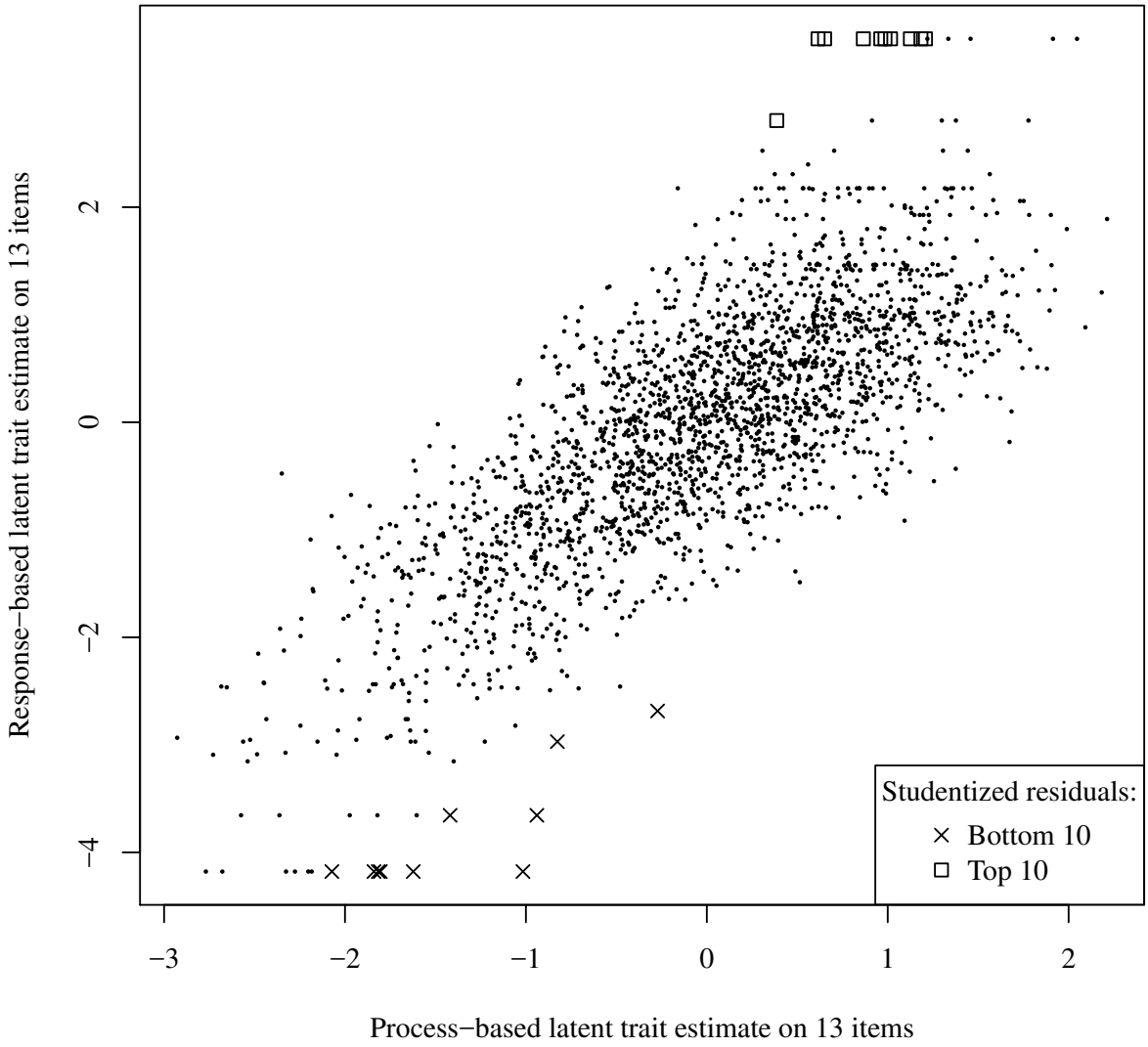

Figure 8: Scatterplot of process- and response-based θ estimates with 13 items (excl. U06a).