Article

Journal of Educational and Behavioral Statistics
2019, Vol. 44, No. 6, pp. 671–705
DOI: 10.3102/1076998619881789
Article reuse guidelines: sagepub.com/journals-permissions

Developments in Psychometric Population Models for Technology-Based Large-Scale Assessments: An Overview of Challenges and Opportunities

Matthias von Davier
Lale Khorramdel
National Board of Medical Examiners

Qiwei He
Hyo Jeong Shin
Haiwen Chen
Educational Testing Service

International large-scale assessments (ILSAs) transitioned from paper-based assessments to computer-based assessments (CBAs) facilitating the use of new item types and more effective data collection tools. This allows implementation of more complex test designs and to collect process and response time (RT) data. These new data types can be used to improve data quality and the accuracy of test scores obtained through latent regression (population) models. However, the move to a CBA also poses challenges for comparability and trend measurement, one of the major goals in ISLAs. We provide an overview of current methods used in ILSAs to examine and assure the comparability of data across different assessment modes and methods that improve the accuracy of test scores by making use of new data types provided by a CBA.

Keywords: *latent regression; process data; mode effects; population model; timing data*

Introduction: New Developments

The use of new technologies in work, education, and everyday life changes the way people think, learn, solve problems, and collaborate. It is important to reflect these new proficiencies and strategies in the assessment frameworks and constructs measured in international large-scale assessments (ILSAs). Consequently, international large-scale studies—such as the Program for the International Assessment of Adult Competencies (PIAAC) and the Program for International Student Assessment (PISA)—moved from a paper-based assessment (PBA) to a computer-based assessment (CBA). This offers new and exciting opportunities

for improvements in data collection, scoring, and analysis but also presents new challenges for analysts and users of the data. CBAs allow for various new interactive item types and the measurement of new constructs or existing constructs with extended frameworks. They also allow for greater efficiency through improved data collection, automated scoring of constructed responses, and the introduction of more complex test designs such as adaptive testing. Moreover, CBA allows the collection of additional information such as response time (RT) and other process-related data (e.g., number of actions, action sequence), which are assumed to be useful for improving the proficiency estimation in item response theory (IRT) scaling and reducing measurement error in latent regressions used in large-scale assessments.

However, changing the mode of administration may affect properties of the tasks and threatens the comparability of data across populations and assessment cycles over time when measuring trends. Both the comparability of data and test scores and the measurement of trends (which is only possible when based on comparable scores) are central to ILSAs. Therefore, possible mode effects need to be evaluated and treated to ensure comparable test scores before any further innovations are introduced. Once comparability and a successful transition from PBA to CBA are established, new information and data provided by the CBA can be used to make further improvements to data analyses and the generation of test scores. Including RT and process data, however, is not straightforward and requires research to evaluate the relation between these new variables and the proficiencies of interest. The interaction of these new variables and their validity has to be investigated before they can be introduced in operational analyses of large-scale assessments. Moreover, there is the need to reduce the large number of distinct variables provided by the CBA into fewer, more meaningful units before including them in further analysis or the population model.

The Latent Regression (Population) Model

In ILSA such as PISA, a latent regression (or population) model is being used to estimate posterior proficiency distributions based on the likelihood function of an IRT model and a linear regression of background data on the proficiency of interest (von Davier, Gonzalez, & Mislevy, 2009; von Davier, Sinharay, Oranje, & Beaton, 2006). It can be viewed as an imputation model for the unobserved proficiency variable that aims to obtain unbiased (or at least less biased) group-level proficiency distributions. This requires the estimation of an IRT measurement model, which provides information about how test performance depends on proficiency, and the latent regression, which provides information about the extent to which background information is related to proficiency. The population model is usually estimated separately for each population of interest (in PISA and PIAAC, this would be the different countries), and a predefined number of plausible values (PVs), which are multiple imputations, are drawn from the

resulting posterior distribution for each respondent (e.g., 10 PV in PISA) in each cognitive domain. It is important to note that PVs are no individual test scores and should only be used for analyses at the group level.

These types of models tend to utilize many variables in the latent regression to avoid missing any useful information collected either from the background questionnaires or from the process data (von Davier et al., 2009). Because of this considerable number of background variables, a principal component analysis (PCA) is used in the latent regression model to reduce the number of variables to a smaller number of meaningful predictors that are accounting for a large proportion of the variation in the background questionnaire variables. For PISA, it was decided to use the components for each country that accounted for 80% of the variance in order to avoid numerical instability due to potential overparameterization of the model (Organization for Economic Cooperation and Development [OECD], 2017). The problem of overparameterization is important with regard to including additional variables in the latent regression model such as process data. This issue will be discussed in our article.

Aim of the Current Article

This article presents an overview of innovations targeting psychometric approaches and methodologies that deal with the comparability of data between modes of administration as well as the new data provided by CBAs and their potential use for improving estimation in ILSAs. The order in which the different sections are presented in the article follows the priority and sequence of the operational implementation in large-scale surveys: (1) The first section deals with using IRT to control mode effects when PBA items are transferred to a CBA in order to establish comparable item parameters and test scores across modes and for the measurement of trends. Only when comparability is established, a population model—which is based on the item parameters obtained in the IRT scaling—can yield comparable test scores across groups, modes, and over time. (2) Once comparability is established, the use of process and timing data provided by the CBA for improving the population model can be explored. Hence, the second section addresses approaches and challenges to incorporating RT data to enhance the validity and comparability of the assessment as well as to improve the item parameter and proficiency estimation. (3) The third section exemplifies the use of process/sequence data models to generate meaningful indicators and predictive features from process data collected in simulation-based tasks. It is discussed how both RT and features generated from process data can potentially be useful in improving the accuracy of population models. (4) As the availability of timing and process data substantially increases the number of covariates in population models used in large-scale assessments, it is imperative to study data reduction strategies. Therefore, the fourth section surveys variable selection approaches to manage the large amount of process

and timing and background data and to support the selection of variables in latent variable regression models to group-level proficiency distributions. This section is presented as the last one because the population model is the final step in analyzing large-scale assessment data by combining results from the IRT model with background information and possibly with new variables and features generated from RT and process data through a population model.

Comparing Data From Different Modalities of Assessment: Modeling Mode Effects

Despite the advantages of computer-based tests, the move from a PBA to a CBA mode poses challenges for the measurement of trend over time because the results of the same test administered in different modes might not be directly comparable. In addition, it has to be established whether comparability of countries' results in large-scale assessments can be maintained if some use different assessment modes (some countries might not be prepared to utilize computers in the assessment, while others had moved to CBA already). Certain items might not function the same across modes and may differ with regard to their difficulty, discrimination, or the composition of skills they tap into. Mode effects may manifest in the form of differential item functioning (DIF) observable on (at least) some of the items when comparing equivalent groups across different assessment modes. This, in turn, can threaten measurement invariance and can cause undesirable changes in comparability of test scores obtained through the population model and the measurement of trend. Extensions of IRT models can be used to test for mode effects and to deal with violations of measurement invariance if effects are present in the data. This section presents an approach to test and control for possible mode effects using the PISA 2015 data.

We argue that as a first step, the types of mode effects that may change the measurement properties have to be established. Second, items with identified mode effects have to be treated in the modeling to provide valid trend measures as well as comparable scores across samples and groups within an assessment cycle. After such an item-level treatment, different subsets of items might present different levels of invariance. The following provides an overview of types of measurement invariance and illustrates approaches and models that can be used to examine mode effects. The models presented here can be used to select an appropriate treatment of items or scores in the final scaling.

Comparability and Measurement Invariance

There are different levels of measurement invariance (Millsap, 2010) that have to be considered before comparing different groups or assessments over time. For valid interpretations of change over time (trend measure), the assessment should ideally exhibit scalar or strong invariance for all items (the same slope and intercept parameters fit the items independent of the mode of

administration) or at least for a large enough subset of trend items while showing weaker forms of invariance for the remaining items (metric invariance where slope parameters are invariant across modes, while intercepts are allowed to be different). The rationale behind this requirement is that trends measured across modalities are expected to be comparable in order to assess change, and trend measures should provide consistent statistical associations across modes, particularly with external variables central to establishing validity. It should be noted that mode effects are just one possible source of measurement invariance. Other sources such as translation errors, technical issues, and language differences have to and are routinely examined and treated as well (e.g., OECD, 2017; Oliveri & von Davier, 2011; von Davier et al., 2006; von Davier & Sinharay, 2014).

Test Design Requirements for Studying Mode Effects

To evaluate the extent to which measurement invariance can be assumed when moving from a PBA to CBA, an appropriate *data collection design* is needed where the same items have been administered in both modes either to the same students in a counterbalanced design or to randomly equivalent groups of students. Quality is ensured by following best practices such as randomization for the study design so that the different modes of delivery can be understood as treatment assignments in an experiment. In order to be able to generalize from such a study, a sufficiently large and representative sample at the level at which inferences are planned is needed. More specifically, if the level of inference is the functioning of tasks in two modes on the international level, the sample must cover the range of abilities that are assessed across countries. If the level of inference is the detection and potential treatment of country-level mode effects, the sample for each country that plans these types of analyses has to be sufficiently large to enable stable estimates of item parameters for country-level inferences. In country-level mode effect studies, this would require that sample sizes to allow stable estimation of item functions for each item in each mode. Typically, this translates into samples that provide thousands of responses per item, depending on the number and type of parameters of the IRT model used, the targeting of the sample relative to the item difficulty, and so on. If the goal is to evaluate items at the international level, across all samples (i.e., at the international level), 100 to 200 responses per item per country may be sufficient, while this would be insufficient for inferences at the country level. In surveys such as PISA where a field trial is used to evaluate mode effects, this would be typically the case (sample sizes in field trials are usually smaller than in main surveys for reasons of cost efficiency).

Once data are available from such an appropriate mode comparison with equivalent groups of randomly assigned respondents, a second step is to compare paper- and computer-based items with respect to their item parameter

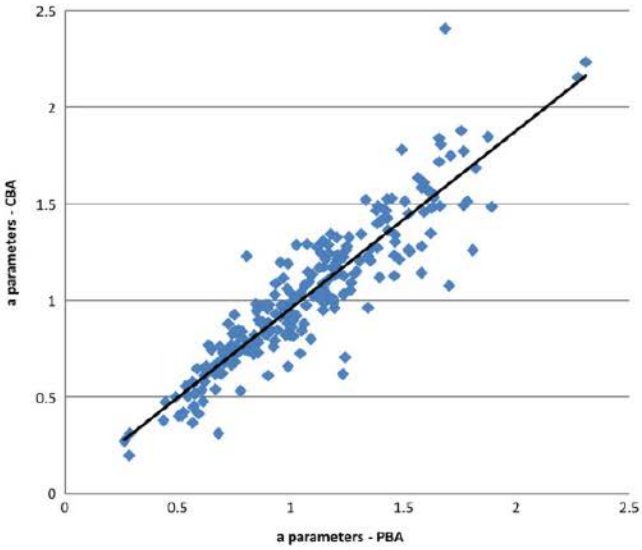

FIGURE 1. *Graphical model check for comparison of slope parameter estimates across paper- (horizontal axis) and computer (vertical axis)-based assessment modes.*

estimates from separate IRT calibrations. In the following, we show how this can be done using graphical model checks and IRT models with different mode effect parameters.

Analysis of Mode Effects Using Graphical Model Checks

As an initial comparison prior to more rigorous modeling approaches, *graphical model checks*, an approach that goes back to Rasch (1960), can be used to spot systematic differences between modes of administration. More information about graphical model checks and their use can be found in Khorramdel and von Davier (2016). Graphical model checks show whether the rank order of item parameters, as well as the relations between item parameters, agrees for all subsamples and thus tests the invariance assumption across subpopulations. Graphical model checks can be used based on data collected in an equivalent groups design. Item parameters are estimated separately for items in paper- and computer-based form while constrained to be equal across country or language groups in order to focus on mode comparisons only and to ensure sufficient sample sizes for calibrations by aggregation over countries.

Figures 1 and 2 show an example of parameter comparisons between modes using the PISA 2015 Field Trial data. The figures contain scatter plots of IRT parameter estimates for different modes.

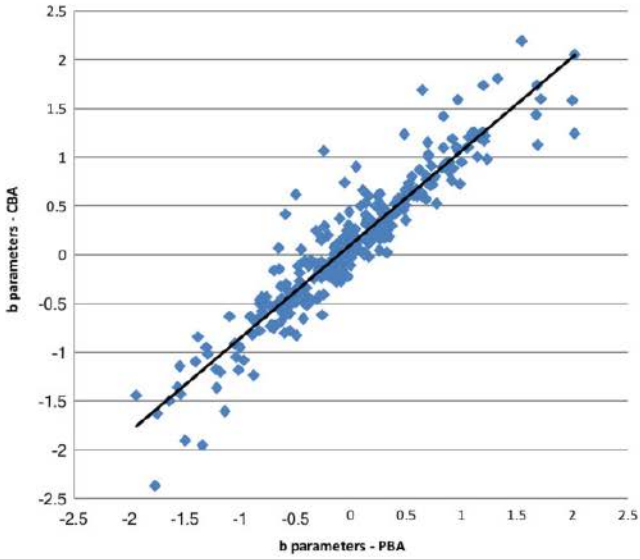

FIGURE 2. *Graphical model check for comparison of difficulty parameter estimates across paper- (horizontal axis) and computer (vertical axis)-based assessment modes.*

Figures 1 and 2 show that item difficulties are highly correlated across modes. The same holds for the freely estimated slope parameters. These results suggest there is good agreement between paper- and computer-based items in terms of retention of difficulty and discrimination across modes. The presence of some outliers, however, suggests that some items differ between modes. These might need treatment to resolve the differences (e.g., splitting the item and estimating separate parameters) in IRT scaling across assessment modes. Note that there is always estimation error in item parameter estimates, especially in the case of aggregate samples using several small field trial samples. Therefore, parameter estimates are not perfectly correlated, and correlations can be particularly low in small within-country samples, mainly due to the sample size but also due to effects of deviations from random assignment and representativeness of small within-country samples. However, correlations at the international level, estimated in the example between item parameters of paper- and computer-based items, are very high, for item difficulty parameters with $r = .94$ and for item slope parameters with $r = .91$. These correlations suggest a statistical link is likely to be established so computer- and paper-based results across countries can be reported on the same scale.

However, before such a link can be established, the extent to which some items may expose mode effects must be examined during IRT scaling. The next section provides an overview of IRT model extensions for examining mode

effects and for linking across assessment modes by testing for, and if present, utilizing the invariance of item parameters across modes.

Mode Effect Models—Accounting for Measurement Invariance and Mode Differences

While graphical model checks can be helpful to examine the overall agreement of item parameters from different samples—for example, tested in different assessment modes—and to explore potential drivers of these differences, they do not provide the most rigorous way to account for mode effects in proficiency estimation (von Davier & von Davier, 2007). In this section, we illustrate how IRT models can be used to analyze mode differences with a higher level of statistical rigor and to achieve unbiased proficiency estimates by treating potential item level effects.

IRT models have been extended to include different types of mode effect parameters in order to provide information about whether the mode effect is best described by an overall difference between assessment modes (i.e., the difference between modes is just adding or subtracting a constant to all assessment tasks), whether it is a person- or group-specific effect that may have an impact differentially on different groups (i.e., some test takers are more affected by mode differences than others), or whether it is a task-specific effect that is only impacting a subset of tasks. These questions can be answered empirically by formalizing these assumptions in a general latent variable model (von Davier, 2008; von Davier, Xu, & Carstensen, 2011) and applying these models to data collected in a randomized mode effect study.

Considering the two-parameter logistic model (Birnbaum, 1968) as the basic model, additional model parameters can be introduced to formalize different assumptions of how mode effects may impact item functioning. Let

$$P(x = 1|\theta, \alpha_i, \beta_i) = \frac{\exp(\alpha_i\theta + \beta_i)}{1 + \exp(\alpha_i\theta + \beta_i)}, \quad (1)$$

denote the probability of a correct response by a respondent with proficiency θ for an item with parameters α_i, β_i. The notation in Equation 1 can be transformed to the customary notation by letting $a = \alpha/1.7$ and $b = -\beta/\alpha$.

Mode Effects on the Item Level

A common (but maybe overly simple) mode effect assumption is that all items are "shifted" by a certain amount with respect to their difficulty when comparing one mode of administration with another. The reason could be that reading or, more generally, processing the item stem or stimulus is generally harder or easier (by the same amount for all items) on the computer or entering a response is more tedious or simpler than bubbling in a response on an answer sheet. A mode effect that homogeneously applies to all items on a test, when changing the mode for all

items, can be modeled by adding the same constant to all difficulty parameters in the case of the affected mode: a general mode effect parameter $-\delta_m$ that represents how much more difficult (or easy) solving any item is when presented in a different mode relative to the Reference Model 1. For items in the "new" model, we assume that

$$P(X = 1|\theta, \alpha_i, \beta_i, \delta_m) = \frac{\exp\left(\alpha_i\theta + \beta_i - 1_{\{I+1,\ldots,2I\}}(i)\delta_m\right)}{1 + \exp\left(\alpha_i\theta + \beta_i - 1_{\{I+1,\ldots,2I\}}(i)\delta_m\right)}. \quad (2)$$

This can be thought of as a model for twice the number of items. The indicator function $1_{\{I+1,\ldots,2I\}}(i)$ equals 1 if the item index is in the second half, that is, the range $I+1, \ldots, 2I$. The first $1, \ldots, I$ items are the paper-based items without mode effect, and the items in the new mode are indexed by $I+1, \ldots, 2I$. In this notation, it is assumed that item i and item $i+I$ are the same but administered in different modes. This leads to a model with $2I$ items (instead of I items for each delivery mode separately) in which the difficulty parameters for items presented in one mode (say, paper) are assumed to be β_i for $i = 1, \ldots, I$ and the item parameters for the other mode (say, computer) are appended as parameters β_j for $j = I+1, \ldots, 2I$ and arranged in the same order and constrained to follow $\beta_{j+1} = \beta_j - \delta_m$. In the equivalent groups design, each test taker receives half of the items, either paper items, indexed by $i = 1, \ldots, I$, or the computer-based items indexed by $i = I+1, \ldots, 2I$.

In contrast to the assumptions of a general mode effect parameter ($-\delta_m$), one could argue that not all items change difficulty when moving from a PBA to CBA: Some could be more difficult, some could be at the same difficulty level, and some could get easier. This leads to a model that adds an item-specific effect $-\delta_{mi}$ to the difficulty parameter. This can be written as a DIF parameter, quantifying item-specific changes from PBA presentation, namely

$$P(X = 1|\theta, \alpha_i, \beta_i, \delta_m) = \frac{\exp\left(\alpha_i\theta + \beta_i - 1_{\{I+1,\ldots,2I\}}(i)\delta_{mi}\right)}{1 + \exp\left(\alpha_i\theta + \beta_i - 1_{\{I+1,\ldots,2I\}}(i)\delta_{mi}\right)}. \quad (3)$$

The difference in comparison to the model of metric (or "weak") factorial invariance (Meredith, 1993) is that the computer-based difficulties that are written in reference to the paper mode are decomposed into two components, that is, $\beta_{i+I} = \beta_i - \delta_{mi}$, while we continue to assume that $\alpha_{i+1} = \alpha_i$ for the slope parameters. This decomposition indicates that the difficulties are shifted by some (item or item feature)-dependent amount, the shift being applied to one mode on an item-by-item basis—one that is being considered the reference mode with no shift. For items that are not significantly affected by mode, we may further impose a model constraint assuming that $\delta_{mi} = 0$.

The model in Equation 3 with constraints across both modes on slope parameters, as well as potential constraints on the DIF parameters, establishes a measurement invariance (e.g., Meredith, 1993) IRT model. This model can be viewed as representing weak factorial invariance. The larger the number of constraints of the type $\delta_{mi} = 0$ can be assumed, the more we approach a model with strong factorial invariance. Note that we already assume the equality of means and variances of the latent variable in both modes because it is assumed that respondents receiving the test in computer or paper mode are randomly selected from a single population.

Mode Effects on the Respondent or Proficiency Level

If it cannot be assumed that the mode effect is a constant (even if item dependent) shift in difficulty for all respondents, one may assume that an additional proficiency (ϑ) is required to accurately model response probabilities for the new mode. This leads to a multidimensional model with a second latent variable that is added to the item function for items administered in the new mode. The expression $\alpha_{mi}\vartheta$ in the model below indicates that there is a second slope parameter α_{mi} for items ($i = I + 1, \ldots, 2I$) administered in the new mode and that the effect of the mode is person dependent and quantified through a second latent variable ϑ. We obtain

$$P(X = 1|\theta, \alpha_i, \alpha_{mi}, \beta_i, \vartheta) = \frac{\exp(\alpha_i\theta + \beta_i - \alpha_{mi}\vartheta)}{1 + \exp(\alpha_i\theta + \beta_i - \alpha_{mi}\vartheta)}. \quad (4)$$

Note that slope parameters (α_i) and item difficulties (β_i) are, as before in Models 2 and 3, equal across modes. However, an additional "mode-slope" parameter (α_{mi}), for $i = I + 1, \ldots, 2I$, needs to be estimated, with constant $\alpha_{mi} = 0$ for $i \leq I$. For the joint distribution $f(\theta, \vartheta)$, assume uncorrelated latent variables, $\text{cov}(\theta, \vartheta) = 0$, to ensure identifiability.

In Equation 4, it is assumed that the effect of the person "mode" variable varies across items, maybe the more plausible variant, but a model with item-invariant effects $\alpha_m\vartheta$ (a Rasch variant of a random mode effect) is feasible. However, an item-specific model is more likely to provide better model data fit. As in Model 3, the link between modes can be viewed as increasingly more invariant the more that slope parameters can be assumed to be $\alpha_{mi} = 0$ for items in the new mode. Each constraint ($\alpha_{mi} = 0$) makes the respective item response functions for items i and $i + I$ identical across modes.

Applied to empirical data, the models defined above can be compared based on overall model selection tools such as the well-known information criteria: Akaike information criterion (AIC), Bayesian information criterion, and "Consistent" AIC (Akaike, 1974; Bozdogan, 1987; Schwarz, 1978, respectively). To provide additional evidence beyond this overall model selection approach, the IRT-based (marginal) reliability of proficiency estimates (Sireci, Thissen, &

Wainer, 1991; Wainer, Bradlow, & Wang, 2007, p. 76) under each model should be examined. The best fitting model should provide a sufficiently high reliability of proficiency estimates. Moreover, it is of interest whether models that show similar model fit also provide similar proficiency estimates. If they do, the more parsimonious model should be preferred. Once a mode effect model that provides the optimal choice in terms of parsimony and fit has been determined, the linkage between modes can be evaluated in terms of the percentage of strictly invariant items showing scalar invariance and the percentage of items that show metric or weak invariance. Moreover, the mode effect parameter can be utilized in subsequent operational IRT scaling and linking calibrations.

Applications of Mode Effect Models

The models presented above were developed to test for mode effects and item invariance across assessment modes. They were used for analysis of the PISA Field Trial data collected in 2014 in preparation of the change from PBA to CBA in the subsequent PISA 2015 Main Survey.

The Mode Effect Models 2 and 3 can be estimated with software that allows multiple group IRT model estimation with parameter constraints on item parameters, and Model 4 can be estimated with software that allows multigroup multidimensional IRT models with parameter constraints. For the PISA 2015 analyses, the software *mdltm* (von Davier, 2005) was used for estimation of the Mode Effect Models 2, 3, and 4 described above.

The application of these models is described in the PISA 2015 technical report (OECD, 2017). All models presented above were vetted by the PISA 2015 Technical Advisory Group for operational use after thorough review and applied to the PISA 2015 Field Trial analysis. These models can be categorized as DIF IRT and bifactor IRT models with parameter constraints. Model identification is easily established as all assume invariance with respect to the main dimension for all parameters, only add additional parameters for mode-related DIF effects (or the second mode-specific dimension), and are estimated using randomly equivalent groups with appropriate equality constraints across populations. In the PISA 2015 Field Trial, randomly equivalent groups were used, and consequently, the proficiency distributions, allowed to be different across countries, could be assumed equal across modes. It was found that Model 3 provides appropriate fit to the data, and mode effect parameters are needed only for some items, while most items showed scalar invariance.

After establishing the type of mode effect, items with mode effects were treated in the final scaling using corresponding model constraints. Items with no mode effects received the same slope and intercept or difficulty parameter across modes (scalar invariance). Items with mode effects received different intercepts across modes (metric invariance). There were no items for which both

difficulty and slope have to be made mode-specific. In summary, this section illustrated that modeling mode effects with IRT models allows:

(1) Identifying the type of mode effect
(2) Identifying items that show a mode effect
(3) Treating items with mode effects in the final IRT scaling

Once the assessment has been successfully transferred from a PBA to a CBA and comparability of item parameters across modes of administration is established, the item parameters can be used in the population model to generate PVs for estimation of group-level results and to examine the relation between the construct of interest and additional variables. The next two sections provide a description of the use of RT and features generated from process data as additional sources of information.

Incorporating RTs as Collateral Information

One additional source of information available through CBA is RT, collected together with item responses. RT typically refers to the time a respondent spends on each item (or certain aspects within the time interval) in an assessment. Interest in RT as representing information about response processes has a long history in psychology, in particular in experimental research of reaction times (e.g., speed-accuracy trade-offs). The literature on RT is extensive and reviewing it is beyond the scope of this paper.[1] Here, we focus on studies where RT are relevant to or have been studied in the context of ILSAs. With the wide availability of technology-based testing, RT data have become much more accessible, and research on RT is getting more attention in the ILSA context. For example, the PISA 2015 public use data provide RT data in milliseconds for each cognitive item and at the scale level for the background questionnaire.

In this section, we describe how RT information can be used to improve item parameter estimation and how RT can be included in the population model directly as additional covariates to possibly improve the modeling of group-level proficiency distributions. Aspects of how comparability across countries can be maintained in ILSAs are also addressed. Moreover, we address practical and theoretical issues that should be considered when RT is used as a source of information in conjunction with item responses.

Improving Item Parameter Estimation by Using RT in Data Quality Analysis

For improving the item parameter estimates, RT can be used in data quality analyses and the scaling process to detect data fabrication and suspicious or unexpected response patterns and to model nonresponse behavior. The use of RT has been investigated in various areas in order to improve validity of the

assessment, including item selection in test design and test assembly, diagnosis on aberrant responding behaviors—such as cheating, rapid guessing, and detection of random responses—and analyses of pacing strategies or differential speededness (e.g., Guo et al., 2016; van der Linden, 2007; van der Linden, Breithaupt, Chuah, & Zang, 2007; van der Linden & Guo, 2008; Wise & Kong, 2005). These applications have shown that RTs are potentially useful variables for statistical analyses of international CBAs.

In particular, for low-stakes assessments such as PISA and PIAAC, RT can be used to monitor respondents' *effort and motivation* with the goal to identify unmotivated responses that bias the data. For example, previous studies (e.g., Wise & Kong, 2005) have suggested using an RT solution behavior (SB) index for each respondent–item combination or an RT effort index for individual respondents based on predetermined thresholds. Meyer (2010) presented a mixture Rasch model with RT components to account for differences in respondents' test-taking behavior. Wise and DeMars (2006) proposed the effort-moderated IRT model by incorporating the SB index, which is similar to the extended HYBRID model for test speededness (Yamamoto & Everson, 1997) except that the classification of behaviors is determined by RT rather than the distribution of responses alone. The authors showed their approach can substantially improve proficiency estimation and item parameter estimation when used in the data cleaning process to exclude problematic responses. For example, when rapid guessing was present, advanced psychometric models that incorporated RT information showed better model fit than the traditional IRT models and yielded more accurate item parameter estimates and proficiency estimates with higher convergent validity through simulation studies (Meyer, 2010; Wise & DeMars, 2006). Moreover, RT can be used to identify *data fabrication* (i.e., faked responses) in ILSAs by examining cases where RT for items may be too short or inconsistent with expected times across different groups or countries (Yamamoto & Lennon, 2018).

In relation to identifying unmotivated or fabricated responses, RT can provide valuable information about item-level *nonresponse behavior*. Particularly, in low-stakes testing, omitted responses may not be missing at random and may need treatment by adding structures such as additional latent variables to the IRT model (e.g., Glas, Pimentel, & Lamers, 2015; Rose, von Davier, & Nagengast, 2016). As mentioned above, establishing thresholds based on RT information can be useful if there is a systematic difference between respondent groups identified on the basis of their RT distribution. As an example, Lee and Jia (2014) found that in the National Assessment of Educational Progress (NAEP), rapid responses are uncorrelated to the underlying latent trait, while responses given after taking some time to process the stimulus show positive correlations with proficiency. In PIAAC, RT thresholds are used to differentiate between omitted responses unrelated to the latent trait (rapid responses) and omitted responses that are related to

the trait and can be treated as incorrect responses in the IRT scaling (Weeks, von Davier, & Yamamoto, 2016).

Comparability of RT Across Countries and the Use of RT in Population Models

Before estimating and applying any statistical model that incorporates RT data in ILSAs, a number of matters should be investigated. First, we examined whether RT data are comparable across countries. Once this is established, we discuss how RT can be incorporated in the population model. For demonstration purposes, we are using the example of PISA 2015. There are two reasons for including RT in the latent regression model on top of the numerous variables included already. First, if RT is available in the public use database (as it is the case in PISA) but not used in the population model, this will result in biased estimates of correlations between proficiency estimates and RT in secondary analyses. Second, RT is informative about how test takers manage their time and help evaluating the validity of responses. Additionally, RT may help classifying respondents into groups that may relate to test-taking strategies and motivation in ILSAs (e.g., Lee & Chen, 2011; Lee & Jia, 2014; Weeks et al., 2016). Therefore, RT can be considered an important covariate of proficiency and task performance that is not only part of the public use database but can potentially contribute to better describe group differences.

When we look at RT data across PISA 2015 countries that took the CBA, distributions of the item-level RT in each domain appeared similar across countries regardless of country's performance (OECD, 2017). Furthermore, examining the distribution of RTs at the item and cluster level, a consistent pattern was observed across countries. However, it was noted that the proportion of fast respondents (who spent less than 5 minutes per cluster or less than 10 minutes on two item clusters) and slow respondents (who used up the maximum allowance of assessment time, i.e., 30 minutes per cluster) vary considerably across countries. This suggests that data cleaning (e.g., censoring or standardization if applicable) and data quality analyses should be conducted at the country level rather than aggregating all countries.

Furthermore, this confirms that country-specific conditional distributions should be used, an approach taken in PISA and PIAAC already for population models. Introducing RT into the population model appears feasible as this prerequisite is already in place. RTs (or functions/aggregates of RTs) can be considered as covariates of proficiency and directly included in the set of predictors used in the population model. Writing the likelihood for a respondent given covariates and responses yields

$$L(\zeta, \Gamma, \Sigma; y_{pg}, x_{pg}) = \int_{\theta} \left(\prod_{i=1}^{I} P(y_{pgi}|\theta; \zeta) \right) \phi(\theta|x_{pg}\Gamma,\ \Sigma) d\theta, \tag{5}$$

with $\prod_{i=1}^{I} P(y_{pgi}|\theta;\zeta)$ denoting the likelihood function of an IRT model and y_{pgi} denoting response (y) to item i by student p in country g. RT variables can be included as a part of covariates x_{pg} along with other background characteristics in the estimation of the latent regression $\phi(\theta|x_{pg}\Gamma,\ \Sigma)$, θ denotes the latent proficiency assuming $\theta_{pg} \sim N(x_{pg}\Gamma,\ \Sigma)$.

In operational analysis, item parameters (ζ) associated with the likelihood function are determined in the item calibration stage, prior to estimation of the population model (von Davier et al., 2006). Alternatively, as proposed by van der Linden, Klein Entink, and Fox (2010), RTs can be used as collateral information at the IRT scaling stage, separately from other student background characteristics but along with the item responses (y). Their approach serves the same goal of improvement in estimation accuracy and reduction of bias.

Necessarily, these population models would be assumed to be country-specific, indicated in Equation 5 by adding an index variable denoting that the conditional distribution of the proficiency given RTs and student background data may vary by country g. This country-specific conditional distribution may have important implications, given that countries may vary in time use and pacing patterns and, hence, in the distribution of RTs and in how RTs may relate to performance. If there was a substantial relationship between RTs and proficiencies, the posterior variances would be reduced, as would be the standard errors associated with overall and subgroup estimates. In addition to this expected gain in estimation, some countries might be interested in analyzing and reporting the results by groups defined in terms of RTs, particularly if RTs can be viewed as proxy for motivation.

Regarding the relationship between RTs and proficiencies, findings from PISA 2015 suggest a substantial relationship between RTs and proficiency: The least proficient students took several minutes less on average to complete item clusters, while more proficient students use more time on each cluster (OECD, 2017). A positive relationship between item difficulty and the correlation between RT and proficiency estimates was consistently observed across countries. An example is given in Figure 3. It was found that for easier items, there is no substantial correlation between item-specific RT and proficiency, but for difficult items, there is a positive correlation between RT and proficiency. High-performing students take more time to respond, while low proficiency is associated with shorter RT. This is in line with recent studies that described the relation between RTs and proficiency as being moderated by item difficulty or task complexity (e.g., Becker, Schmitz, Goritz, & Spinath, 2016; Dodonova & Dodonov, 2013; Goldhammer, Naumann, & Greiff, 2015). Results presented here are based on a mixed format test (the figure shows empty dots for binary items and solid dots for polytomous items) of the PISA 2015 Science domain and might be specific to tests with a wider range of item formats and item difficulties, while tests with fewer item types and more similar difficulties may show more

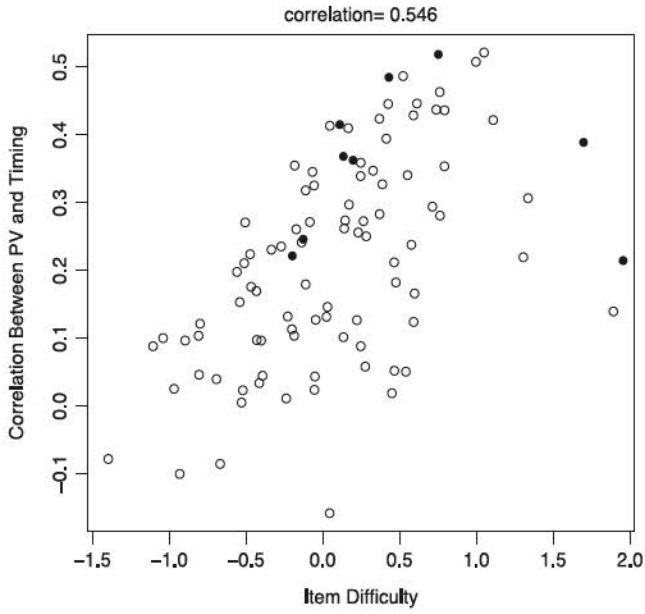

FIGURE 3. *Relationship between item difficulty and correlation between proficiency and response times (one example country); empty dots represent binary items and solid dots represent polytomous items.*

homogeneous associations between the correlation of RTs and proficiency and item difficulty. Also, further investigation is needed with respect to working speed because this relationship is based on the assumption that students work with constant speed. Fox and Marianti (2016) found a similar pattern and partly explained this relationship by allowing differential working speed across the items.

While these preliminary findings support incorporating RTs in the population modeling, more research is needed on how to best include RT information in the conditioning model and whether it can ultimately contribute to improving the estimation of proficiency distributions. More specifically, there are various ways to handle and preprocess the RT data, such as application of PCA to the RTs similarly for student background predictors, using aggregate or summary statistics such as means or medians of the RTs, or generating a typology of students based on their RTs (e.g., students who used the maximum time allowed, students who quit the test quickly, or assigning latent class membership based on respondents' time use patterns). Further investigations are needed whether RTs on individual items can be directly included in the set of predictors for a PCA or should be preprocessed separately from other background characteristics. Also, there is some indication that RTs are multidimensional from our preliminary

findings, which calls for a more thorough investigation of the nature of RTs collected in ILSAs. A recent pilot study using the PISA 2015 data incorporated RT of cognitive items in the population model as an aggregated variable that can be viewed as a proxy of working speed (Shin, Khorramdel, von Davier, Robin, & Yamamoto, 2018). To reduce item-by-person interactions and to include RT as a person covariate, RT was included as a categorized variable and standardized at the country level (i.e., within each country), the item level and the item cluster level. It was found that the inclusion of RT, in contrast to a model without RT, increased the measurement accuracy at extreme proficiency levels (decreased residual variance) without impacting the posterior mean and standard deviation at the country level. However, further research is needed as results varied across countries, and only a subset of countries was examined.

Issues in RT Modeling and Future Research

In order to use RT in ILSAs as additional background data in population modeling, more research is needed, and practical issues raised in the data cleaning process should be carefully considered and explicitly evaluated. For instance, the way RT is preprocessed should be improved with respect to detection and treatment of outliers as well as the transformation and standardization of the RT. As an example, outliers that are censored may still have an impact on estimation, while removing outliers case-wise or replacing outliers by missing values (or expected values or more sophisticated methods of imputations) will pose a different type of issue. Moreover, standardization of RT by item, item types, or cluster positions may impact the results as well. In fact, when the respondents in PISA 2015 were grouped based on a Gaussian finite mixture models via expectation–maximization algorithm (Fraley & Raftery, 2002), the level of standardization (i.e., standardization by item at the country level vs. standardization by item at the test-form level within a country) led to different numbers of clusters in the optimal solution (Shin & von Davier, 2017). As the application of different data cleaning rules can bring about differential impacts in secondary analyses, it is important to examine the possible practices before incorporating the timing data in population modeling in order to improve the accuracy of the proficiency estimation.

In terms of theoretical issues, there are several topics that need more thorough investigation and research in the context of ILSAs. First, we have limited understanding of the nature of RT, in particular, the dimensionality and the distribution of RT collected in low-stakes ILSAs. As noted above, most of the literature has treated RT as a unidimensional entity, while RTs could be multidimensional, implying that RT for an item confounds problem-solving speed with the construct-irrelevant factors such as tendency to guess, motivation, or time management skills (e.g., a choice of how much time to persist on the item or when to stop attempting to solve an item). In addition, the distribution of RT is quite data

dependent. One of the popular choices for the distribution of RT is a lognormal distribution (e.g., van der Linden, 2007). However, the Weibull distribution (e.g., Rouder, Sun, Speckman, Lu, & Zhou, 2003) or γ distribution (e.g., Maris, 1993) appears to be more successful in terms of model–data fit in other cases.

Second, in most of the widely used psychometric RT models, RT and corresponding responses are assumed to be conditionally independent (e.g., van der Linden, 2007). However, the dependency between proficiency and RT may not be as simple or uniform as shown in Figure 3 (i.e., moderated by task difficulty). Respondents are likely to answer items with varying speed over the assessment (e.g., differential working speed such as students work faster at the end of assessment when the time pressure is increased), and individual respondents may take different test-taking strategies and utilize time accordingly. Several recent studies investigated such possibilities by focusing on the interaction between proficiency and RT beyond the linear relation (e.g., Molenaar, Tuerlinckx, & van der Maas, 2015; Partchev & De Boeck, 2012) or by modeling varying speed with respect to the relationship with proficiency (e.g., Fox & Marianti, 2016; Molenaar, Oberski, Vermunt, & De Boeck, 2016). For methods that model responses and RT jointly, model parameter estimation relies on the relationship between speed and accuracy, and RT becomes part of the respondent's scores. Thus, it is important to understand the nature of RT collected in ILSA context to establish the comparability across cycles and countries and the validity of the scores and model parameters before they are of practical use.

Using Process and Sequence Data in Population Modeling

Computer-based testing in ILSAs has made greater data collection flexibility possible, including the capability to administer dynamic and interactive problems, engage students more fully, and capture more information about the problem-solving process (OECD, 2014). Next to RT, a variety of process data such as action sequences can be recorded in log files accompanying test performance data when students respond to items. These data not only help disclose how students solve tasks but also provide collateral information about the response process in addition to the traditional pattern of responses. The availability of process data holds the promise for advancing the science of large-scale assessment by enabling researchers to explore potential reasons for students' success and failure on certain item types and by potentially improving the reliability of the measurement, for example, by providing evidence that may help decide how to treat item-level nonresponse. This can be helpful as part of the data quality evaluation and for the data cleaning process in the operational part of ILSAs when data files are prepared for the public use. Generating and selecting fewer meaningful variables from the large amount of available log-file data facilitates the use of process data in research and enables the potential inclusion of log-file data in the population modeling of ILSAs for improving accuracy and validity.

Validity and Comparability

The extraction of meaningful variables or features from sequence data (Kudenko & Hirsh, 1998) is an essential task in log-file data analyses across items and groups of respondents. Evidence for the validity of process data, that is, the relation of features extracted to one another and to proficiency estimates, must be examined thoroughly before process data can be included in any operational ILSA analysis. Moreover, the comparability of the features and their relation to other variables across different groups and countries has to be investigated.

Unlike features defined in natural language processing (NLP), features of sequences in general process data are not explicit and need to be defined either by experts or extracted using machine learning techniques (Dong & Jian, 2007; Xing, Pei, & Keogh, 2010). In large-scale surveys such as PISA and PIAAC, studies on process data have been conducted in which features were found to be predictive of success and failure. These studies were carried out using data from selected countries and were limited to items that were released by OECD for examination (e.g., He & von Davier, 2015, 2016; Goldhammer et al., 2014; Greiff, Wüstenberg, & Avvisati, 2015). He and von Davier (2016) used process data from PIAAC studying how action sequences from problem-solving tasks are related to task performance. They extracted features from action sequences and found that actions related to using software tools such as sorting and searching occurred significantly more often in the group of respondents that produced a correct response, while actions suggesting hesitation such as repeatedly clicking the cancel button were found more often in respondents giving incorrect responses. In comparisons across countries, it was found that action sequences significantly differed by performance groups but were consistent across countries. The consistently most predictive ("robust") features extracted from a combined sample that distinguished success and failure groups were consistent with those extracted separately within each country.

Studies regarding student strategies in a simulation-based environment provide evidence on the comparability of process data across items in similar settings. Greiff, Wüstenberg, and Avvisati (2015) investigated students' problem-solving strategies in an interactive item from PISA 2012. Experts in scientific inquiry suggested that students who use a strategy of varying one thing at a time (VOTAT) are more likely to succeed. It turned out that this feature has the highest correlation with success on the task. A similar result was found by Han, He, and von Davier (2016), where VOTAT ranked as one of the top predictors of students' success along with features generated using machine-learning techniques. However, since all described studies were done on a single or a few selected items only, more studies are needed on a broader set of tasks.

Feature Generation and Selection From Process Data

Feature generation is defined here as a process to create variables based on timing and process data either by aggregating sequence data or by detecting patterns in sequences captures in log files. An example of a simple aggregate would be the frequency of a specific action in an action sequence (e.g., "how often did the test taker press the cancel button"). The generated features can be roughly categorized into three groups: (1) behaviors that represent respondents' problem-solving strategies manifested as interactions with the computer, as in the strategy indicator VOTAT; (2) actions and mini-action sequences (e.g., *n*-grams that disassemble the long action sequences into small action sections that contains one-, two-, or three-adjacent actions) that are directly extracted from test takers' process data; and (3) timing data, such as time spent on the test, time spent in the simulation environment, and time to the first action when solving a task. The input of content experts and item developers is valuable when developing sequence features that are believed to be associated with the problem-solving process. More specifically, experts might be able to formalize what they expect to see in a proficient respondent with regard to how they approach a complex interactive problem.

Feature selection models are helpful to identify consistently the most predictive ("robust") indicators that distinguish different performance groups. A variety of models have been developed in fields now often described as "big data" applications for feature selection. For large-scale assessment sequence data, He and von Davier (2016) used the chi-square (χ^2) selection model (CHI; Oakes, Gaizauskas, Fowkes, Jonsson, & Beaulieu, 2001) to extract predictive actions and action sequences from process data in a pilot study using data from PIAAC. CHI is recommended for use in textual analysis due to its high effectiveness in finding robust key words and for testing similarities between different text corpora (e.g., He, Glas, Kosinski, Stillwell, & Veldkamp, 2014; He, Veldkamp, & de Vries, 2012; He, Veldkamp, Glas, & de Vries, 2017; Manning & Schütze, 1999, for more feature selection models, refer to Forman, 2003). Because of the structural similarity between text and process data, it appears appropriate to apply this approach to detect actions or action vectors that are highly informative for distinguishing performance groups.

The χ^2 feature selection takes the basic premise of the traditional χ^2 test by comparing the frequencies of events in a 2×2 contingency table as shown in Table 1. The (weighted) number of action occurrences in two groups, C_1 (i.e., correct group) and C_2 (i.e., incorrect group), is indicated by n_i and m_i, respectively. The sum of the weighted action occurrences in each group is defined as the group length $len(C)$. The idea behind this method is to test whether occurrence and nonoccurrence of actions and correctness of the item response are independent. Thus, the method compares two groups to determine how far C_1 deviates from C_2 in terms of action frequencies.

TABLE 1.
2 × 2 Contingency Table for Action i *in* χ^2 *Score Calculation*

	C_1	C_2
action i	n_i	m_i
¬ action i	$len(C_1) - n_i$	$len(C_2) - m_i$

Note. C_1 and C_2 represent the two study groups (e.g., correct and incorrect), n_i and m_i indicate the weighted frequency of the action i occurs in C_1 and C_2, respectively, and $len(C)$ is the sum of the weighted action occurrences in each group.

Note that term weights are taken into account when calculating the χ^2 score for each action sequence. Analogous to the weighting scheme in NLP, the weighting scheme could consist of two parts: (1) inverse sequence frequency (ISF) of an action i as $\text{ISF}_i = \log(N/\text{sf}_i) \geq 0$, where N indicates the total number of sequences in the collection, namely, the total number of test takers and sf_i is the number of sequences where the action i appears. This weight helps eliminating low-informative (ubiquitous) actions and favors highly informative (rare) actions. (2) Another concern about term frequency is regarding clustering at the individual level. The importance of an action that is taken multiple times by one individual should be different from that when the action is taken once each by multiple individuals. We dampen the term frequency by a function $f(\text{tf}) = 1 + \log(\text{tf})$, $\text{tf} > 0$ because more occurrences of a word indicate higher importance, but not as much relative importance as the undampened count would suggest.

These two parts could be further combined into a single weight as follows:

$$\text{weight}(i,j) = \begin{cases} [1 + \log(\text{tf}_{i,j})]\log(N/\text{sf}_i) & \text{if } \text{tf}_{ij} \geq 1 \\ 0 & \text{if } \text{tf}_{ij} = 0 \end{cases}, \tag{6}$$

where N is the total number of sequences. The first clause applies to actions occurring in the same sequence, whereas for not observed actions ($\text{tf}_{i,j} = 0$), we use $\text{weight}(i,j) = 0$.

Under the null hypothesis, the two collections of action sequences are randomly equivalent, so the distribution of actions should not differ between correct and incorrect groups. A χ^2 value is computed to evaluate the departure from this null hypothesis. For a 2×2 contingency table, the χ^2 value is computed as

$$\chi^2 = \frac{M(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})}, \tag{7}$$

where M is the total number of actions in the collection and O_{ij} represents the weighted counts in each cell in the matrix (Agresti, 1990; Bishop, Fienberg, & Holland, 1975). Actions with higher χ^2 scores are more discriminative in classification (Manning & Schütze, 1999). Therefore, we ranked the χ^2 score of each action in descending order. The actions ranked at the top were defined as robust classifiers.

He and von Davier (2016) selected the action features by different performance groups (correct/incorrect) on an aggregate level and a country level. Consistency of action sequence patterns was found across countries. The features extracted by the χ^2 model in a selected problem-solving item were compared with those selected by the weighted log likelihood ratio (WLLR) test in He and von Davier (2015). The WLLR is defined as the product of the probability of each action sequence and the logarithm of the ratio between the conditional probabilities of the sequence in different performance groups. Analogous to the WLLR being applied in feature selection for text categorization (Nigam, McCallum, Thurn, & Mitchell, 2000), the WLLR of each action sequence for the purpose of binary classification (i.e., correct group and incorrect group at the item level) can be defined as follows:

$$\mathrm{WLLR}(a, C_i) = p(a|C_i)\log\frac{p(a|C_i)}{p(a|\bar{C}_i)}, \tag{8}$$

where $p(a|C_i)$ is the conditional probability of action a given in group C_i and $p(a|\bar{C}_i)$ is the conditional probability of action a in the complement (respondents not in group C_i). This comparison study not only validated the previously identified actions but also showed that the mini sequences (e.g., bigrams and trigrams) were more useful than single actions (i.e., unigrams) in describing test takers' behaviors in the process data analysis.

Han et al. (2016) applied random forests (RF; e.g., Breiman, 2001; Dietterich, 2000) to extract 15 predictors from 77 features generated from process data that were collected on a problem-solving item (climate control) in PISA 2012. The extracted 15 predictors yielded 85.7% accuracy in classifying the correct and incorrect response group, which is only marginally lower than using the total of 77 features (which yields 89.2% accuracy). The 77 features consist of action sequences at the aggregate level (e.g., the order of using simulation settings) and action level (n-grams or mini-sequences of respondent actions), strategy indicators (e.g., VOTAT), and timing features (e.g., problem-solving total time). The 15 selected features appear to be sufficient to describe student strategies in that they allow a prediction of task performance at a level of accuracy that is very close to the level reached when using all features. One advantage of using RF for feature selection is that it can deal with highly correlated features and multiple interactions without resorting to simple aggregates of actions that may not have a clear interpretation. Another advantage is that RF allows utilizing best practices that can be viewed as RF-specific versions of well-known statistical tools such as cross-validation and resampling.

Integration of Process Data Into Population Models

Process data play an important role in validating response data and supporting the interpretation of test takers' performance. Moreover, they have the potential

to play a key role in the population modeling of computer-based large-scale assessments, assuming the model can be extended to include these variables, that is, selected features from log file data including behavioral indicators, actions and mini-sequences, and timing data. However, some challenges in using process data for population modeling need to be confronted. In particular, process data indicators are typically collected only at the level of the tasks in which they occur; higher level aggregation into variables that cut across tasks are the exception. There are at least two challenges for incorporating process data into large-scale population models: (1) the sparseness and large proportion of missing data and (2) the large total numbers of predictors:

(1) Incomplete block designs are used in all major survey assessments. As a consequence, cognitive indicators as well as process indicators are available only for the tasks a student has taken, while the observations on these variables are missing for all other tasks for this student in the incomplete block design. Such an incomplete design is not favorable for population modeling (von Davier, 2013) as it makes strong conditional independence assumptions (Little & Rubin, 2002).

(2) The sheer number of potential process data variables is very large. Due to the fact that only a subset may be observed per person, there needs to be some level of aggregation across items or clusters. However, even with aggregated variables such as those described in the feature generation and selection processes above, the number of additional indicators to be included in the population model is potentially very large and may effectively make the application of standard regression techniques impossible, as the number of variables relative to the number of cases in the sample may be too large.

The next section provides a somewhat selective overview of models and approaches that may help overcome these challenges.

Variable Selection Methods for Using Timing and Process Data in Population Models

Like any regression-based model, the population model potentially suffers from an issue that too many predictor variables may be used in the model, leading to over parameterization. The sections above discussed how this already large pool of variables used in operational analyses may be further extended by the availability of timing and process data in CBAs. This section provides an overview of selected models and approaches that may help overcome these challenges. In particular, approaches that can help select variables in prediction models as well as approaches that allow dealing with missingness by design will be discussed. There are certainly more methods available than can be described here due to length constraints.

In current operational use of the population model in assessments such as NAEP, PISA, Trends in International Mathematics and Science Study, and Progress in International Reading Literacy Study, a PCA is used to preprocess the

predictors. This is done with the goal of reducing the number of variables and to remove collinearity, so that the latent regression can be estimated more stably. Usually, principal components (PCs) are used that accounted for 80% or 90% of the variance to avoid numerical instability due to potential overparameterization of the model. Using PCs serves to retain information for students with missing responses to one or more background variables. However, if the main aim is to include background variables that correlate with the proficiency variable, it appears likely that not all background variables are needed as predictors, as most of the predictive power likely comes from a small number of variables (Thomas, 2002). While prior assessments may provide some insight into which background variables are relevant, it remains unclear whether new types of variables obtained from CBAs such as timing and process indicators can add predictive power to the latent regression. The sheer amount of process data collected in CBAs makes it necessary to utilize methods that help selecting a relevant set of variables to come up with a set of predictors that is optimal for the purpose of generating public use files with PVs.

Several methods have been developed to find an optimal set of predictor variables in regression models. However, not all methods may be directly applicable in the case of latent regression models. In this section, we describe a number of these approaches and compare their benefits and shortcomings, especially for a population model that may contain timing and process data along with more traditional covariates collected in background questionnaires.

First Studies on Variable Selection in Population Models

The first studies on variable selection in population models are based on data from NAEP as this was the first large-scale assessment program to use the latent regression population model (Mislevy, 1991) as well as PCs extracted from the background questionnaire data for the estimation of the latent regression parameters. Mazzeo, Johnson, Bowker, and Fong (1992) first hint at the need to select variables in the population model based on an R^2 criterion in order to reduce the number of parameters in the latent regression. An analysis by Kaplan and Nelson (see Mislevy, 1991) using the 1988 NAEP reading data suggested that a small number of the PCs will capture most of the proficiency variance and produce almost identical proficiency distributions while reducing the chance of overfitting the model. Thomas (2002) presents similar results and provides evidence that the use of a large number of PCs that explain 90% of the variance of the background questionnaire variables adds little over the use of a small number of major reporting variables (gender, ethnicity, limited English proficiency, individualized educational plan) only.

Moreover, Thomas (2002) examined whether the accuracy of the latent regression model (with regard to the recovery of the proficiency distribution) can be maintained by using only a small number of primary reporting variables

mentioned above or whether the accuracy of a model that uses secondary (auxiliary) variables as predictors in addition to primary reporting variables is superior. The results of this study indicate that improvement in precision depends on the matrix sampling design used for the cognitive assessment. The improvement in precision observed when including more (secondary/auxiliary) variables range from essentially none for a balanced design (different respondents may receive different sets of items, which are linked, but each respondent is asked to answer items in all domains) to moderate for a split design where not all respondents receive items for all constructs of interest. Based on his findings, Thomas suggests to reduce the collection of covariates and increase the number of cognitive items administered to each respondent (in each domain) and to eliminate the use of (secondary) covariates from the creation of PVs. These findings suggest that the selection of variables and covariates in latent regression models is complex and need thorough investigation. An alternative approach to increasing the number of cognitive items and potentially loosing information by reducing the use of secondary variables could be the additional use of RT and process data. Process variables are closely related to the cognitive items and can provide additional information about the latent trait of interest. However, this would lead to an increase in predictor variables, which makes the need for variable selection methods and tools more obvious. A few possible methods are described in the following.

Model-Building Algorithms for Variable Selection

An interesting area to learn from is regression modeling that is aimed at machine learning applications. These models are based on vast amounts of data with a (very) large number p of predictors $X_1, \ldots, X_p$. For linear regression models, there are several model-building algorithms used for variable selection such as forward selection, backward elimination, all subsets regression, and various combinations (Efron, Hastie, Johnstone, & Tibshrani, 2004). The aim is to produce "good" linear models for predicting a response y on the basis of some selected covariates $x_1, x_2, \ldots, x_p$. Parsimonious or simpler models having only a small number of nonzero parameters are preferred over models with many parameters for the sake of numerical stability and ease of interpretation (Hastie, Tibshirani, & Wainwright, 2015). Goodness of fit for model selection is often defined and evaluated in terms of prediction accuracy. In the following, we provide a short overview of a few selected approaches.

Forward Selection

Given a collection of possible predictors, forward selection (or forward stepwise selection) is used to select the one predictor that has the largest absolute correlation with the response y, say, x_{j1}; then, a simple linear regression of y on

x_{j1} is performed (Weisberg, 1980). This leaves a residual vector orthogonal to x_{j1}, which is now considered to be the response. This selection process is repeated with other predictors. After k steps, this results in a set of predictors $x_{j1}, x_{j2}, \ldots, x_{jk}$ that are then used in the usual way to construct a k-parameter linear model. The problem of this method is that it is overly greedy, perhaps eliminating some useful predictors at the second step that happen to be correlated with x_{j1}.

Forward Stagewise and Least Absolute Shrinkage and Selection Operator (LASSO)

The forward stagewise approach (e.g., Efron et al., 2004) is an iterative process that begins with the regression parameters of β as 0 and builds up the regression function in successive small steps. If β is the current stagewise estimate, let $c(\beta)$ be the vector of current correlations

$$c(\beta) = X'(y - X\beta). \tag{9}$$

Then, we take a very small step in the direction of the forward (stepwise) selection, based on the direction and size of the correlations. Since the size of the step is very small, it takes many more steps to get the final result compared with forward (stepwise) selection.

Tibshirani (1996) developed the LASSO approach, which is very similar to the forward stagewise algorithm. Let $x_1, x_2, \ldots, x_p$ be vectors representing the covariates, and let y be the vector of responses for the n cases with the assumptions that the covariates have been standardized to have mean = 0 and unit length and that the response has mean = 0. The LASSO can be viewed as a penalized regression approach where

$$\hat{\beta}_{\text{LASSO}} = \arg\min_{\beta} \left\{ \frac{1}{N} ||y - X\beta||_2^2 \right\}, \text{ with } ||\beta||_1 \leq t, \tag{10}$$

is the estimator that minimizes the mean squared error subject to a constraint, which is predetermined by the user. The larger the t, the bigger size of the β. If one chooses t to be ∞, then the forward stagewise becomes the forward (stepwise) selection. The constraint produces a regression in which several regression parameters differ from 0 (as the contribution of each parameter to the penalty term is positive whenever $\beta_j \neq 0$). More information about LASSO and other methods that exploit sparsity to help recover the underlying signal in a set of data can be found in Hastie, Tibshirani, and Wainwright (2015).

In a simulation study, Hastie, Taylor, Tibshirani, and Watlher (2007) compared the forward stagewise algorithm with the LASSO. Despite their findings that LASSO is less constrained and allows sudden changes of direction where forward stagewise (which behaves like a monotone version of LASSO) tends to slow down the search, they conclude that forward stagewise might be preferable

for models with a large number of predictors. The reason is that in their study, the coefficient paths for forward stagewise showed to be smoother while those for LASSO fluctuated widely (due to strong correlations of subsets of variables). Moreover, in later stages, forward stagewise takes longer to overfit, likely due to the smoother paths.

Least Angle Regression (LARS)

LARS is similar to the stagewise procedure using a mathematical formula to accelerate computations. The LARS procedure starts with all coefficients equal to zero and finds the predictor most correlated with the response, say, x_{j1}. The largest step possible in the direction of this predictor is taken until some other predictor, say, x_{j2}, has as much correlation with the current residual. LARS proceeds in a direction *equiangular* between the two predictors until a third variable x_{j3} earns its way into the "most correlated" set. LARS then proceeds in a direction equiangular between x_{j1}, x_{j2} and x_{j3}, that is, along the "least angle direction," until a fourth variable enters, and so on. Therefore, LARS will only need m steps to find a solution; here, m is the number of all predictors. Under certain conditions or modifications, the LARS algorithm can yield all forward stagewise or LASSO solutions. Selecting rather conservative criteria within LARS (or other approaches) might be useful to avoid the problem of missing important predictors.

Background Data Reduction Using Latent Class Analysis (LCA)

LCA-based methods are another possible approach for reducing the number of predictors in population models. LCAs can identify one or more latent nominal variables that can be used to classify respondents with respect to their background characteristics. These classifications can then be introduced as predictors in the population model. Wetzel, Xu, and von Davier (2015) compared different LCAs to the more traditional PCA approach and showed a reduction in predictor variables in the population model when using dummy-coded maximum a posteriori LCA-based class membership indicators as predictors. The recovery of the group means and standard deviations of the operational approach (PCA) was quite satisfactory for all examined LCA models. Furthermore, the posterior means and standard deviations used to generate PVs derived from the PCA and the LCA approaches were very similar.

The advantages of the LCA methods are a more meaningful explanation from the group background classes than the PCs of the PCA can offer and the use of cases with missing data in background variables. Moreover, the LCA approach does not require all the background variables to be contrast-coded, and in general, fewer coefficients need to be estimated compared with the PC approach.

Conclusion and Outlook

Most major ILSAs have transitioned or are in the process of transitioning from a PBA to a CBA. PIAAC introduced a CBA in addition to a PBA in 2012, PISA moved to a CBA for most participating countries in 2015, and in 2016, NAEP was piloting the administration of the assessment on tablets. This new computerized digital form of testing offers new opportunities but also poses new challenges. On one hand, CBAs allow for the collection of RT and process or log-file data that can be used to improve data quality and subsequent data cleaning processes and could be included as additional predictor variables in latent regression population models to improve the accuracy of group-level proficiency measures. On the other hand, the transition from PBA to CBA poses a potential problem for comparability and the measurement of trends, which are both important goals in ILSAs. Possible changes in item characteristics may occur (some items might be easier in one mode than in the other, for example), which threatens measurement invariance and need to be accounted for in data analysis. Therefore, the first step in analyzing data from an ILSA which just moved to a CBA is to examine and treat possible mode effects for providing comparable item parameters and stable trend measures, even before new types of data can be introduced in data analysis procedures and population modeling. Not only the comparability of test scores obtained from population models is challenged, even numerous data quality analyses rely on the comparability of item parameters and trend comparisons (checking whether a country performs differently on trend items in two different assessment cycles is one example). Once measurement invariance and comparability of item parameters across modes of administration are established, new statistical methods to include RT and process data can be examined in a next step to possibly improve operational ILSA analyses.

This article describes modeling approaches needed for the challenges of the transition from a PBA to a CBA and for the use of RT and process data with a special focus on improving population models in ILSAs. First, the models required for a transition to CBA, aimed at controlling for mode effects, are discussed. We, thereby, focus on IRT-based modeling approaches as used for operational analyses in ILSAs (an alternative Bayesian approach to examine violations of measurement invariance is presented by Verhagen and Fox, 2012, which enables multiple marginal invariance hypotheses to be tested simultaneously; this seems to be a promising development but has, so far, not been implemented in operational ILSA analyses, to our knowledge). Second, once a successful transition is achieved, modeling approaches are described that allow us to better utilize and select the additional information gained by the new assessment mode. The collection of RT and process data in addition to the test takers' responses on items—simply put, all actions taken by a test taker are stored and saved in log files—can potentially be useful for test developers, psychometricians, and researchers but leaves us with an enormous amount of data and with

databases with very large number of variables. One of the challenges is how to use these augmented databases in meaningful ways for research and how to include the information contained in process data in population models for providing more accurate and fair test scores. How can timing and process data be analyzed and related to proficiency? How can large numbers of variables, not all of which are available for all respondents due to incomplete designs, be transformed into fewer meaningful aggregates and how can variables be selected for further analyses? The current article addresses these questions and gives an overview of some relevant approaches that may guide the way and show us how to use this information in scaling and population modeling. Finally, reporting in large-scale assessments relies on collateral and background data, which are used in population models to add information to the cognitive items for more accurate group-level proficiency measures. Since population models must handle a very large number of background variables already, additional information from timing and process data add to the challenge of possible overparameterization of a modeling approach that is central to reporting and database generation. Therefore, the process of variable selection in population modeling has to be improved, and new approaches have to be examined and compared to find an optimal balance. In this article, we give an overview of selected approaches for illustration purposes without the aim to cover the wide range of available methods and algorithms.

The current article does not only illustrate and introduce recent approaches to deal with new information from CBAs and big data but seeks to stress the importance of using them carefully and the need for further research. What is needed are validity studies for log-file data and improvements of the single statistical approaches before they can be used to generate official group-level assessment results (proficiency scores and PVs) or for secondary analyses.

Authors' Note

This is an invited article for the *JEBS* special issue on innovations in large-scale assessment methodology.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

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

Note

1. Readers can refer to a recent special issue in the *British Journal of Mathematical and Statistical Psychology* (2017, volume 70, issue 2) for some of the state-of-the-art RT models.

References

Agresti, A. (1990). *Categorical data analysis*. New York, NY: Wiley.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Becker, N., Schmitz, F., Göritz, A. S., & Spinath, F. M. (2016). Sometimes more is better, and sometimes less is better: Task complexity moderates the response time accuracy correlation. *Journal of Intelligence*, *4*, 11.

Birnbaum, A. (1968). *On the estimation of mental ability* (Series Report No. 15). Randolph Air Force Base, TX: USAF School of Aviation Medicine.

Bishop, Y. M., Fienberg, S., & Holland, P. (1975). *Discrete multivariable analysis*. Cambridge, MA: MIT.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*, 345–370.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. doi:10.1023/A:1010933404324

Dietterich, T. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, *40*, 139–157.

Dodonova, Y. A., & Dodonov, Y. S. (2013). Faster on easy items, more accurate on difficult ones: Cognitive ability and performance on a task of varying difficulty. *Intelligence*, *41*, 1–10.

Dong, G., & Jian, P. (2007). *Sequence data mining*. New York, NY: Springer.

Efron, B., Hastie, T., Johnstone, I., & Tibshiran, R. (2004). Least angle regression. *Annals of Statistics 2004*, *32*, 407–499.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, *3*, 1289–1305.

Fox, J. P., & Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multivariate Behavioral Research*, *51*, 540–553.

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, *97*, 611–631.

Glas, C. A. W., Pimentel, J. L., & Lamers, S. M. A. (2015). Nonignorable data in IRT models: Polytomous responses and response propensity models with covariates. *Psychological Test and Assessment Modeling*, *57*, 523–541.

Goldhammer, F., Naumann, J., & Greiff, S. (2015). More is not always better: The relation between item response and item response time in Raven's matrices. *Journal of Intelligence*, *3*, 21–40.

Goldhammer, F., Naumann, J., Selter, A., Toth, K., Rolke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, *106*, 608–626.

Greiff, S., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers & Education*, *91*, 92–105.

Guo, H., Rios, J. A., Haberman, S., Liu, O. L., Wang, J., & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education*, *29*, 173–183.

Han, Z., He, Q., & von Davier, M. (2016). *Predictive feature generation and selection from process data in simulation-based environment: An implementation of Random Forest*. Paper presented at the 78th National Council on Measurement in Education (NCME) Conference, Washington, DC.

Hastie, T., Taylor, J., Tibshirani, R., & Watlher, G. (2007). Forward stagewise regression and the monotone LASSO. *Electronic Journal of Statistics*, *1*, 1–29.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: The LASSO and generalizations*. Boca Raton, FL: CRC Press.

He, Q., Glas, C. A. W., Kosinski, M., Stillwell, D. J., & Veldkamp, B. P. (2014). Predicting self-monitoring skills using textual posts on Facebook. *Computers in Human Behavior*, *33*, 69–78.

He, Q., Veldkamp, B. P., & de Vries, T. (2012). Screening for posttraumatic stress disorder using verbal features in self narratives: A text mining approach. *Psychiatry Research*, *198*, 441–447.

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

He, Q., & von Davier, M. (2015). Identifying feature sequences from process data in problem-solving items with n-grams. In A. van der Ark, D. Bolt, S. Chow, J. Douglas, & W. Wang (Eds.), *Quantitative psychology research: Proceedings of the 79th annual meeting of the psychometric society* (pp. 173–190). New York, NY: Springer.

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. 749–776). Hershey, PA: Information Science Reference.

Khorramdel, L., & von Davier, M. (2016). Item response theory as a framework for test construction. In K. Schweizer & C. Distefano (Eds.), *Principles and methods of test construction: Standards and recent advancements* (pp. 52–80). Göttingen, Germany: Hogrefe.

Kudenko, D., & Hirsh, H. (1998). Feature generation for sequence categorization. In *AAAI '98/IAAI '98: Proceedings of the 15th National/10th Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence* (pp. 733–738). Menlo Park, CA: American Association for Artificial Intelligence.

Lee, Y.-H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling*, *53*, 359–379.

Lee, Y.-H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-Scale Assessments in Education*, *2*. doi:10.1186/s40536-014-0008-1

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York, NY: Wiley.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Maris, E. (1993). Additive and multiplicative models for gamma distributed variables, and their application as psychometric models for response times. *Psychometrika*, *58*, 445–469.

Mazzeo, J., Johnson, E., Bowker, D., & Fong, Y. F. (1992). *The use of collateral information in proficiency estimation for the trial state assessment*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA, April 20–24.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525–543.

Meyer, J. P. (2010). A mixture Rasch model with item response time components. *Applied Psychological Measurement*, *34*, 521–538.

Millsap, R. E. (2010). Testing measurement invariance using item response theory in longitudinal data: An introduction. *Child Development Perspectives*, *4*, 5–9. doi:10.1111/j.1750-8606.2009.00109

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*, 177–196.

Molenaar, D., Oberski, D., Vermunt, J., & De Boeck, P. (2016). Hidden Markov IRT models for responses and response times. *Multivariate Behavioral Research*, *51*, 606–626.

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. (2015). A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times. *Multivariate Behavioral Research*, *50*, 56–74.

Nigam, K., McCallum, A. K., Thurn, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, *39*, 103–134.

Oakes, M., Gaizauskas, R., Fowkes, H., Jonsson, W. A. V., & Beaulieu, M. (2001). A method based on chi-square test for document classification. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 440–441). New York, NY: ACM.

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling*, *53*, 315–333. Retrieved from http://www.psychologie-aktuell.com/fileadmin/download/ptam/3-2011_20110927/04_Oliveri.pdf

Organization for Economic Cooperation and Development. (2014). *PISA 2012 results: Creative problem solving: Students' skills in tackling real-life problems* (Vol. V). Paris, France: Author. doi:10.1787/9789264208070-en

Organization for Economic Cooperation and Development. (2017). *PISA 2015 technical report*. Paris, France: Author. Retrieved from http://www.oecd.org/pisa/sitedocument/PISA-2015-Technical-Report-Chapter-9-Scaling-PISA-Data.pdf

Partchev, I., & De Boeck, P. (2012). Can fast and slow intelligence be differentiated? *Intelligence*, *40*, 23–32.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Nielsen & Lydiche (Expanded edition, Chicago, University of Chicago Press, 1980).

Rose, N., von Davier, M., & Nagengast, B. (2016). Modeling omitted and not-reached items in IRT models. *Psychometrika*. doi:10.1007/s11336-016-9544-7

Rouder, J. N., Sun, D., Speckman, P. L., Lu, J., & Zhou, D. (2003). A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika*, *68*, 589–606.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Shin, H. J., Khorramdel, L., von Davier, M., Robin, F., & Yamamoto, K. (2018). *Incorporating response time into population modeling for large-scale assessments*. Paper presented at the 11th Conference of the International Test Commission (ITC), Montreal, Canada.

Shin, H., & von Davier, M. (2017). *Understanding time usage patterns and their associations with proficiencies in international large-scale assessments*. Paper presented at the Timing Impact on Measurement in Education (TIME) conference held by National Board of Medical Examiners, Philadelphia, PA.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*, 237–247.

Thomas, N. (2002). The role of secondary covariates when estimating latent trait population distributions. *Psychometrika*, *67*, 33–48.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of Royal Statistical Society. Series B*, *58*, 267–288.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287–308.

van der Linden, W. J., Breithaupt, K., Chuah, S. C., & Zhang, Y. (2007). Detecting differential speededness in multistage testing. *Journal of Educational Measurement*, *44*, 117–130.

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*, 365–384.

van der Linden, W. J., Klein Entink, R. H., & Fox, J.-P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement*, *34*, 327–347.

Verhagen, A. J., & Fox, J. P. (2012). Bayesian tests of measurement invariance. *British Journal of Mathematical and Statistical Psychology*, *66*, 383–401.

von Davier, M. (2005). *A general diagnostic model applied to language testing data*. (ETS Research Report No. RR-05-16). Princeton, NJ: Educational Testing Service.

von Davier, M. (2008). The mixture general diagnostic model. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 255–274). Charlotte, NC: Information Age.

von Davier, M. (2013). Imputing proficiency data under planned missingness in population models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*. Boca Raton, FL: CRC Press (Chapman & Hall).

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful? In *IERI Monograph Series: Issues and Methodologies in Large Scale Assessments*, *Vol. 2*. Retrieved from IERI website: http://www.ierinstitute.org/IERI_Monograph_Volume_02_Chapter_01.pdf

von Davier, M., & Sinharay, S. (2014). Analytics in international large-scale assessments: Item response theory and population models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.). *Handbook international large-scale assessment: Background, technical issues, and methods of data analysis*. Boca Raton, FL: CRC Press (Chapman & Hall).

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (2006). Statistical procedures used in the National Assessment of Educational Progress (NAEP): Recent developments

and future directions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics (Vol. 26): Psychometrics* (pp. 1039–1055). Amsterdam, the Netherlands: Elsevier.

von Davier, M., & von Davier, A. A. (2007). A unified approach to IRT scale linking and scale transformations. *Methodology*, *3*, 115–124.

von Davier, M., Xu, X., & Carstensen, C. H. (2011). Measuring growth in a longitudinal large-scale assessment with a general latent variable model. *Psychometrika*, *76*, 318–336.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York, NY: Cambridge University Press.

Weeks, J., von Davier, M., & Yamamoto, K. (2016). Using response time data to inform the coding of omitted responses. *Psychological Test and Assessment Modeling*, *58*, 671–701.

Weisberg, S. (1980). *Applied linear regression*. New York, NY: Wiley.

Wetzel, E., Xu, X., & von Davier, M. (2015). An alternative way to model population ability distributions in large-scale educational surveys. *Educational and Psychological Measurement 2015*, *75*, 739–763.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*, 19–38.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*, 163–183.

Xing, Z., Pei, J., & Keogh, E. (2010). A brief survey on sequence classification. *SIGKDD Explorations Newsletter*, *12*, 40–48.

Yamamoto, K., & Everson, H. (1997). Modeling the effects of test length and test time on parameter estimation using the hybrid model. In J. Rost (Ed.), *Applications of latent trait and latent class models in the social sciences* (pp. 89–98). Münster, Germany: Waxmann.

Yamamoto, K., & Lennon, M. L. (2018). Understanding and detecting data fabrication in large-scale assessments. *Quality Assurance in Education*, *26*, 196–212.

Authors

MATTHIAS VON DAVIER holds the Distinguished Research Scientist position at National Board of Medical Examiners, 3750 Market Street, Philadelphia, PA 19104; email: mvondavier@nbme.org. His research interests are quantitative psychology, educational measurement, large scale assessments, deep learning, and computational statistics.

LALE KHORRAMDEL is a senior research scientist at National Board of Medical Examiners, 3750 Market Street, Philadelphia, PA 19104; email: lkhorramdel@nbme.org. Her research interests are psychometrics, large scale assessments, and response styles.

QIWEI HE is a research scientist at Educational Testing Service, 660 Rosedale Road, Princeton, NJ 08540; email: qhe@ets.org. Her research interests are psychometrics and process data analysis.

HYO JEONG SHIN is an associate research scientist at Educational Testing Service, 660 Rosedale Road, Princeton, NJ 08540; email: hshin@ets.org. Her research interests psychometrics, rater modeling, and timing data analyses.

HAIWEN CHEN is a senior research scientist at Educational Testing Service, 660 Rosedale Road, Princeton, NJ 08540; email: hchen@ets.org. His research interests are psychometrics and software development.

Manuscript received April 14, 2017
First revision received February 7, 2018
Second revision received March 13, 2019
Accepted September 20, 2019