PSYCHOMETRIKA—VOL. 80, NO. 3, 727–747
SEPTEMBER 2015
DOI: 10.1007/s11336-014-9415-z

PREDICTIVE INFERENCE USING LATENT VARIABLES WITH COVARIATES

LYNNE STEUERLE SCHOFIELD

SWARTHMORE COLLEGE

BRIAN JUNKER

CARNEGIE MELLON UNIVERSITY

LOWELL J. TAYLOR

CARNEGIE MELLON UNIVERSITY

DAN A. BLACK

UNIVERSITY OF CHICAGO

Plausible values (PVs) are a standard multiple imputation tool for analysis of large education survey data, which measures latent proficiency variables. When latent proficiency is the dependent variable, we reconsider the standard institutionally generated PV methodology and find it applies with greater generality than shown previously. When latent proficiency is an independent variable, we show that the standard institutional PV methodology produces biased inference because the institutional conditioning model places restrictions on the form of the secondary analysts' model. We offer an alternative approach that avoids these biases based on the mixed effects structural equations model of Schofield (Modeling measurement error when using cognitive test scores in social science research. Doctoral dissertation. Department of Statistics and Heinz College of Public Policy. Pittsburgh, PA: Carnegie Mellon University, 2008).

Key words: latent variable analysis, NAEP, plausible value methodology, marginal estimation procedures..

1. Introduction

Latent variable models for measuring cognitive constructs (e.g., proficiency in a particular domain of mathematics) are ubiquitous in education research and institutional reporting. Item response theory (IRT) models (van der Linden & Hambleton, 1997; Fox, 2010) offer the machinery needed to handle sophisticated item- and person-sampling schemes in complex survey data. Even in simpler settings these models offer proficiency estimates with high reliability and precision, due to their efficient use of assessment data. Econometricians, policy analysts, and other social scientists increasingly rely on the results of latent variable measurement models to generate constructs for statistical analysis.

Studies that characterize students' achievement under different curricula, compare students belonging to different social groups, or evaluate achievement differences across countries, use estimated proficiency as the *dependent variable* in their analyses. Studies that focus on downstream outcomes, such as earnings in the labor market, might assess the direct effect of academic proficiency on the outcome, or control for proficiency in trying to assess the influence of other

Correspondence should be made to Lynne Steuerle Schofield, Department of Mathematics and Statistics, Swarthmore College, 500 College Avenue, Swarthmore, PA 19081, USA. E-mail: lschofi1@swarthmore.edu

variables of interest. In these latter cases, estimated proficiency is an *independent variable* in the analysis.

Whether latent proficiency variables play the role of dependent or independent variables, the issue of measurement error must be addressed. If the proficiency variables were estimated without error, they could be used directly with no adjustments. If, as is usually the case, proficiencies are estimated with some uncertainty, it will affect both the precision and bias of estimated effects. The accurate assessment of precision requires using appropriately adjusted standard errors or similar calculations. Bias must be dealt with by conditioning proficiency estimates on an appropriate set of covariates.

The appropriate conditioning model has been discussed at length by Mislevy (1991), Mislevy, Beaton, Kaplan, and Sheehan (1992), and others. A key motivating application is the release of data for secondary analysis by large institutional surveys, such as the U.S. National Assessment of Educational Progress (NAEP, Allen, Carlson, & Zelenak, 1999), or other large-scale national and international surveys of education that have a similar structure (e.g., the National Adult Literacy Survey, NALS, (Kirsch et al., 2000); and Trends in International Mathematics and Science Study, TIMSS, Olson, Martin, & Mullis, 2008). In a data release that provides individual-level proficiency measures, a fixed number of multiple imputations (Rubin, 1987, 1996) for each individual's proficiency are released. These imputations, known as *plausible values* (PVs) in this context, are adjusted to account for degraded precision and bias due to measurement error, in two ways. First, they are Monte Carlo draws from posterior proficiency distributions for each individual, and hence incorporate all sources of uncertainty (including measurement error). Second, the posterior distribution is conditioned not only on the individual responses to items on a cognitive assessment, but also on a set of demographic and other background variables. PV methodology provided in Mislevy (1991), Mislevy et al. (1992), and other sources allow secondary analysts to account for measurement error in subsequent analyses by employing PVs in appropriate ways (e.g., Mislevy, 1991, 1993; von Davier, Gonzalez, & Mislevy, 2009). Typically, agencies release five PVs for each individual, along with instructions for using PVs to estimate regression coefficients and other effects. (For more on current PV methodology, see Li, Orange, & Jiang, 2009.)

Given standard practice, there is a subtle but important question about the conditioning model used to generate PVs: What data, aside from the item response data themselves, should be incorporated in generating the posterior distribution from which PVs are drawn? Based on an argument developed by Mislevy (1991), institutions that release PVs typically condition on a fixed but extremely large set of covariates to account for the large universe of studies that a secondary analyst might undertake. In particular, any contrast (such as a comparison between mean proficiencies in two social groups of interest) that a hypothetical secondary analyst might be interested in must be included, directly or by proxy, in the conditioning model used by the institution to generate PVs. In Sects. 2 and 3, we review this argument and see that when proficiency is a *dependent* variable the release of institutional PVs based on an extremely large conditioning model allows a secondary analyst to conduct estimation that is unbiased but perhaps statistically inefficient, under rather general assumptions. In Sect. 4, we provide a disquieting result for the case when proficiency is an *independent* variable in a regression model. We show that secondary analysis is susceptible to substantial bias when using institutional PVs as independent variables with the standard methodology prescribed for them. Because of the complex nature of the conditioning model, a secondary analyst has essentially no chance of specifying a model consistent with the survey institution's modeling choices. The bias involves not only the regression coefficient for proficiency in the model, but also regression coefficients for other predictors, whether they are latent or not. Our conditions and results are similar in spirit to Meng's (1994) work on *congeniality* between inference methods and imputation models, but our work is distinct from Meng's in two ways: first, we consider imputation for latent variables, which are measured with some error, and second we consider the role of latent variables as independent variables in secondary analyses.

An immediate consequence of these results is that the use of institutional PVs based on a large, fixed conditioning model may introduce substantial bias when proficiency is an independent variable in a secondary analysis. More broadly, analysts who wish to use latent variables to predict other outcomes should use conditioning models that are customized to their particular prediction problem; in Sect. 5 we discuss workable machinery to do this.

Our results are stated in considerable generality. Nevertheless, we show an example in Sect. 6 to demonstrate the size and direction of the bias when the secondary analyst's model is not compatible with the institution's conditioning model. In the example, the structural model of interest is a linear regression model in which proficiency serves as an independent variable predicting weekly wages, and the measurement model is a standard IRT model. We use data from the 1992 National Adult Literacy Survey (NALS) to illustrate the bias resulting from the incompatibility of the secondary analyst's structural model and the institution's conditioning model.

2. Modeling Components for the Analysis of Education Surveys

Before discussing the key results of Mislevy (1991) and Mislevy et al. (1992) in Sect.3 and exploring their extension to models that use proficiency to predict other outcomes in Sect. 4, it is important to describe and discuss the two sets of analysis that modern large-scale education surveys are designed to serve, and to discuss, in abstract terms, the tools that they use to make inferences from the survey data.

In order to focus the discussion on these two sets of analyses and their inferential tools, we will ignore some complexities of education surveys, such as complex student-sampling designs (which are generally accounted for with design-based survey weights and jackknife or Taylor linearization variance adjustments), and complex item-sampling designs (which are generally accounted for with incomplete likelihoods in the measurement model, e.g., items not administered are missing completely at random or MCAR by design). All of these complexities, and the tools developed to address them, are crucial for practical inference from education survey data, and are well described in the technical documentation for these surveys (e.g., Allen et al., 1999; Allen, Donoghue, & Schoeps, 2001; Kirsch et al., 2000; NCES, 2009; Olson, Martin, & Mullis, 2008). But to review them in detail would distract from the essential structure of the inferential problems faced both by *primary analysts* working on behalf of the *survey institution*, and by *secondary analysts* who use public use and restricted use data to answer questions not envisioned in the reports published by the survey institution.

In Sect. 2.1, we review the survey institution's *measurement model*, a generative psychometric model that describes the relationships between participants' proficiency in a particular cognitive domain, and their responses to particular cognitive items in the survey. In Sect. 2.2, we review the survey institution's *population model*, also known as the *conditioning model*, which describes variation in cognitive proficiency across groups defined by demographic, jurisdictional, and other background covariates. We then briefly discuss in Sect. 2.3 the kinds of inference made by primary analysts working on behalf of the survey institution. Finally, in Sect. 2.4, we discuss two basic classes of models used by secondary analysts to explore research questions not contemplated in the survey institution's reports.

2.1. The Measurement Model

For simplicity, we suppose there are N participants (students or other respondents) in the education survey and J test items. We denote the response of participant i to question j as X_{ij}, and the set of all responses as $X = [X_{ij}]_{i=1,j=1}^{N,J}$. We also denote the latent proficiency of the ith

participant as θ_i, and the set of all N proficiencies as $\theta = (\theta_1, \ldots, \theta_N)$. The *measurement model*

$$p(X|\theta) \tag{1}$$

is the generative model chosen by the survey institution to model the likelihood of observing response matrix X, given latent proficiencies θ. The measurement model may depend on other parameters, which do not concern our analysis here.

The formulation in (1) is intended to be quite general and cover a broad variety of possible stochastic models for measurement, including

- classical unidimensional dichotomous item response theory (IRT) models, which take the form

$$p(X|\theta) = \prod_{i=1}^{N}\prod_{j=1}^{J} P(\theta_i|\gamma_j)^{X_{ij}}(1 - P(\theta_i|\gamma_j))^{1-X_{ij}},$$

  where $X_{ij} = 0$ or 1, indicating a wrong or right response, θ_i is a single real number indicating a level of proficiency, $P(\theta)$ is a standard item characteristic curve, such as the 2-parameter logistic (2PL) model, and γ_j is a set of item parameters for item j, such as a discrimination parameter a_j and difficulty parameter b_j, in which case $\gamma_j = (a_j, b_j)$;
- multidimensional IRT (MIRT) models, in which θ_i is a vector of d real numbers, $\theta_i = (\theta_{1i}, \ldots, \theta_{di})$, denoting proficiencies on d latent constructs, and $P(\theta)$ and γ_j are modified accordingly;
- polytomous variations on the IRT or MIRT models above, in which X_{ij} can take values in a discrete set of categories, and $P(\theta)$ and γ_j are modified accordingly;
- cognitive diagnosis models (CDMs), in which X_{ij} may take dichotomous or polytomous values, θ_i is a d-dimensional vector of discrete indicators denoting the presence or absence of d specific skills or knowledge components, and $P(\theta)$ and γ_j are modified to specify a specific CDM such as the DINA or DINO model;
- factor analysis (FA) models in which X_{ij} are continuous responses, θ_i is a vector of continuous factor scores, and $p(X|\theta)$ is a typical FA model; and
- other models in which X_{ij} is a more complex (multivariate, graphical, etc.) response, and/or θ_i is a more complex proficiency variable, and/or the measurement model $p(X|\theta)$ may reflect violations of, or variations on, many standard assumptions such as local independence, monotonicity, experimental independence, complete data, etc.

In most modern large-scale education surveys, the measurement model (1) is some form of an IRT or MIRT model. Whatever the measurement model is, it is usually pre-calibrated so that any item parameters $\gamma_1, \ldots, \gamma_J$ can be thought of as fixed and known for all subsequent analyses. We will assume this in our development below. In the case in which γ_j are estimated along with other quantities, however, there is no essential change in the message of our work.

2.2. *The Population (Conditioning) Model*

In a typical large-scale education survey, the survey institution is primarily interested in reporting on features of the *population distribution*

$$p_{\mathrm{PA}}(\theta|Z). \tag{2}$$

The subscript PA in (2) is intended to remind that this distribution is the focus of the *primary analysis* performed by the survey institution or its contractors. The variable Z denotes an entire set of conditioning variables that are of interest in the primary analysis. These might typically include

- primary reporting variables;
- survey design variables;
- jurisdictional or institutional variables that describe the institutions (typically schools, school districts, governmental jurisdictions, etc.) that the individual participants (typically students) are members of;
- variables concerning participants' education contexts, such as teacher questionnaires;
- participant demographic variables, such as gender, race/ethnicity, age, SES; and
- other background variables for individual participants, such as education experience or number of hours of TV watched, which might be collected through a background questionnaire administered to individual participants.

The conditioning variables Z in our setup subsume both the collateral variables Y and the design variables Z in the setup of Mislevy (1991) and Mislevy et al. (1992). The distinction between design variables and collateral variables is not important for our development, and we wish to reserve Y for the (observable) dependent variable in a prediction model.

The model in (2) is highly multivariate in both θ and Z. Indeed, Z generally spans "reporting variables" that serve primary analyses and reports by the survey institution, other demographic, background, and jurisdictional variables that may serve secondary analysts, and many interactions between them. Thus, (2) usually conditions on a large set of covariates Z, and so it is also known as the *conditioning model* for the survey.

2.3. *Primary Analysis: Reporting and Plausible Values*

It is typically not possible to do inference on small jurisdictional units or individual participants, for a variety of legal and technical reasons. Many education surveys, such as the National Assessment of Educational Progress, operate under laws that proscribe the public identification of individual students, schools, or other local units that participate in the survey. More broadly, it is generally unethical to break confidentiality and privacy commitments made to survey participants. At a more technical level, the participant sample is usually not designed to provide reliable inferences at the level of a school or even a school district of moderate size (and would be prohibitively expensive if it were so designed), and the number of cognitive items asked of any individual participant is small enough (to manage time and fatigue constraints) that inference for an individual is usually not reliable either. Instead, the targets of inference for primary analysis by the survey institution are typically means, percentiles, and other summaries for major reporting groups, defined by reporting variables such as race/ethnicity, gender, age, region, larger jurisdictions, as well as some background variables.

The "institutional model" described by Eqs. (1) and (2) is essentially a two-stage generative model for the cognitive data collected in the survey: first, θ is generated from its conditional distribution given Z, and then X is generated from its conditional distribution given θ. The objects of inference for primary analyses are features of the θ distribution, after collecting the survey data. This suggests a Bayesian, or at least empirical Bayes, approach. Indeed, the measurement model in (1) can be though of as a likelihood for θ, and the conditional model in (2) can be thought of as a prior distribution for θ. Then the posterior distribution of θ may be written as

$$\begin{aligned} p_{\mathrm{PA}}(\theta|X, Z) &\propto p(X|\theta, Z)\, p_{\mathrm{PA}}(\theta|Z) \\ &= p(X|\theta)\, p_{\mathrm{PA}}(\theta|Z) \end{aligned} \tag{3}$$

under the assumption that $X \perp\!\!\!\perp Z \mid \theta$, which is usually true by design of the measurement process producing X (i.e., if the measurement process is well designed, X should be conditionally independent of any other variable, given θ). In typical settings, a great deal of X is missing by design, to reduce testing time, fatigue, etc., for individual participants. The mechanics of implementation of the measurement model (1), as reported in the technical documentation for any large-scale education survey—such as Allen et al. (1999), (2001), Kirsch et al. (2000), NCES (2009), and Olson et al. (2008)—allow reporting for all groups of students on a common θ scale. Thus, different groups are equated on a common θ scale, even though they may have seen disjoint sets of items.

The summaries (e.g., conditional means or percentiles) produced in primary reports by the survey institution are either *functionals* of the posterior distribution $p_{PA}(\theta|X, Z)$—that is, they can be obtained by computing the integral[1]

$$\int s(\theta, Z)\, p_{\mathrm{PA}}(\theta|X, Z) d\theta \tag{4}$$

for some appropriate function $s(\theta, Z)$—or they can be derived from functionals of $p_{PA}(\theta|X, Z)$. The quantities in (3) and (4) may be estimated using Bayesian methods (Johnson & Jenkins, 2005) or marginal maximum likelihood and other methods (Allen et al., 2001).

Following the work of Mislevy (1991), Mislevy et al. (1992), and others, many survey institutions compute and publish *plausible values* (PVs) for θ in large-scale education surveys. PVs, known in the statistics literature as multiple imputations (Rubin, 1996), are sets of random draws from the posterior distribution (3). Their primary use, as noted by Mislevy et al. (1992, p. 142), is as a Monte Carlo numerical integration tool for integrals such as (4). PVs and their appropriate use have been discussed recently by von Davier et al. (2009), and the consequences of their misuse in certain contexts was recently discussed by Carstens and Hastedt (2010).

2.4. *The Secondary Analyst's Research Models*

In addition to primary reports generated by the survey institution and its contractors, important substantive and methodological work has been performed by *secondary analysts* (NCES, 2008; Robitaille & Beaton, 2002), that is, researchers acting independently of the survey institution, investigating questions outside the scope of the primary reports. Substantive questions for secondary analysts often revolve around some feature of a distribution such as

$$p_{\mathrm{SA}}(\theta|\tilde{Z})\ , \tag{5}$$

where the subscript SA is intended as a reminder that this is a model chosen by the secondary analyst and $\tilde{Z}$ represents covariates of interest to the secondary analyst in his/her research. For example, if the components of $\theta = (\theta_1, \ldots, \theta_N)$ are continuous and unidimensional, and $\tilde{Z}$ can be separated into participant-level pieces $\tilde{Z} = (\tilde{Z}_1, \ldots, \tilde{Z}_N)$, then $p_{\mathrm{SA}}(\theta|\tilde{Z})$ might be expressed as a linear model

$$\theta_i = \beta_0 + \beta_1 \tilde{Z}_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2). \tag{6}$$

In general $\tilde{Z}_i$ need not be univariate, in which case β_1 is a vector of regression coefficients. In addition $\tilde{Z}$ may or may not be identical to Z in (2). Indeed, the most interesting and innovative

[1] As suggested in Sects. 2.1 and 2.2, X, Z and θ are extremely general multidimensional objects; they may have components that are continuous, discrete, etc. For ease of exposition, we will express all appropriate probability calculations as integrals, as if the variables involved were continuous. For other variable types, the integrals can be replaced with appropriate sums, Riemann–Stieltjes integrals, etc., as needed. The essential message of our work is the same.

secondary analyses usually involve $\tilde{Z}$ not contemplated by the survey institution. Because θ appears as the dependent variable in the regression form of (5), we refer to models of the form (5) as θ-*dependent* models.

The object of inference in the θ-dependent case is typically some function $s(\theta, \tilde{Z})$, which captures some feature of $p_{\mathrm{SA}}(\theta|\tilde{Z})$ of interest. For example, in the linear regression case, the secondary analyst might be interested in the least-squares estimate of β_1, in which case $s(\theta, \tilde{Z})$ takes the form

$$s(\theta, \tilde{Z}) = \hat{\beta}_1 = \frac{\widehat{\mathrm{Cov}}(\theta, \tilde{Z})}{\widehat{\mathrm{Var}}(\tilde{Z})},$$

where $\widehat{\mathrm{Cov}}(\cdot, \cdot)$ and $\widehat{\mathrm{Var}}(\cdot)$ denote sample covariance and variance calculations suitable for the design of the survey. More generally, the posterior distribution of β_1,

$$s(\theta, \tilde{Z}) = p(\beta_1|\theta, \tilde{Z}),$$

and similar quantities, may be of interest.

Another class of models considered by secondary analysts, especially those interested in using cognitive proficiency in predicting later outcomes Y, is of the form

$$p_{\mathrm{SA}}(Y|\theta, \tilde{Z}). \tag{7}$$

Under suitable assumptions, this might also be expressible as a regression model such as

$$Y_i = \beta_0 + \beta_1\theta_i + \beta_2\tilde{Z}_i + \varepsilon_i\,, \quad \varepsilon_i \sim N(0, \sigma^2)\,, \tag{8}$$

where, once again, $\tilde{Z}$ may or may not be identical to Z, and either or both of θ_i and $\tilde{Z}_i$ might be multidimensional. Since θ appears as an independent variable in the regression form of (7), we refer to models of the form (7) as θ-*independent* models.

The object of inference in the θ-independent case is again some function $s(\theta, Y, \tilde{Z})$, which now captures some feature of $p_{\mathrm{SA}}(Y|\theta, \tilde{Z})$ of interest. For example, in the linear regression case, the secondary analyst will typically be interested in the least-squares estimate of some regression coefficient(s) or the posterior distribution of the regression coefficient(s).

3. The θ-Dependent Case

We consider first the θ-dependent case. Here, the secondary analyst has a research model of the form of (5), i.e.

$$p_{\mathrm{SA}}(\theta|\tilde{Z}),$$

in which θ appears as the dependent variable, and the object of inference is a function $s(\theta, \tilde{Z})$ related to $p_{\mathrm{SA}}(\theta|\tilde{Z})$. Because θ is completely missing, the most the secondary analyst can hope to learn is some feature of a marginal quantity such as

$$s(X, \tilde{Z}) = \int s(\theta, \tilde{Z})\, p_{\mathrm{SA}}(\theta|X, \tilde{Z})\mathrm{d}\theta = E_{\mathrm{SA}}[s(\theta, \tilde{Z})|X, \tilde{Z}]. \tag{9}$$

We can now state, in modern terms, the problem identified and solved by Mislevy (1991) and Mislevy et al.(1992): What tool can the primary analyst provide to the secondary analyst, so that (a) an integral of the form of (9) can be calculated or approximated appropriately, and (b) the results of the secondary analysis are numerically consistent with the primary survey reports? The answer is the publication of institutional PVs, as discussed above at the end of Sect. 2.3.

With institutional PVs the secondary analyst can approximate the quantity

$$s(X, \tilde{Z}, Z) = \int s(\theta, \tilde{Z}) p_{\mathrm{PA}}(\theta|X, Z)\mathrm{d}\theta = E_{\mathrm{PA}}[s(\theta, \tilde{Z})|X, Z]\,, \tag{10}$$

which is a functional of the form (4). The following theorem lays out conditions under which calculation of (10) leads to an unbiased estimate of $s(X, \tilde{Z})$, providing the underlying justification for the use of PVs as outlined by Mislevy (1991) and subsequent authors.

Theorem 3.1. *If $\tilde{Z} \subseteq Z$, then $s(X, \tilde{Z}, Z)$ is an unbiased estimate of $s(X, \tilde{Z})$.*

By the notation $\tilde{Z} \subseteq Z$, we mean that the σ-field generated by $\tilde{Z}$ is a subfield of the σ-field generated by Z (see Billingsley, 1986, for definitions). Informally, this means that $\tilde{Z}$ is a deterministic function of Z.

Proof. We calculate

$$\begin{aligned} E_{\mathrm{SA}}\{s(X, \tilde{Z}, Z)|X, \tilde{Z}\} &= E_{\mathrm{SA}}\{E_{\mathrm{PA}}[s(\theta, \tilde{Z})|X, Z]|X, \tilde{Z}\} \\ &= E_{\mathrm{SA}}[s(\theta, \tilde{Z})|X, \tilde{Z}] \\ &= s(X, \tilde{Z}) \end{aligned}$$

by the "telescoping" property of conditional expected values, when $\tilde{Z} \subseteq Z$ (Billingsley, 1986, p. 470). □

Biases arising when $\tilde{Z} \not\subseteq Z$ have been illustrated by Mislevy et al. (1992), von Davier et al. (2009) and Carstens and Hastedt (2010).

The standard procedure for using institutional PV's, described by Mislevy (1991, pp. 181–182), amounts to calculating

$$s_m = s(\theta_m, \tilde{Z}), \quad m = 1, \ldots, M$$

for each of M imputations θ_m drawn from $p_{\mathrm{PA}}(\theta|X, Z)$, and then averaging. This produces

$$\bar{s} = \frac{1}{M}\sum_{1}^{M} s_m \approx \int s(\theta, \tilde{Z}) p_{\mathrm{PA}}(\theta|X, Z)\mathrm{d}\theta = E_{\mathrm{PA}}[s(\theta, \tilde{Z})|X, Z],$$

the Monte Carlo numerical approximation to $s(X, \tilde{Z}, Z)$ in (10). A further between/within variance calculation (Mislevy, 1991, p. 182) approximates the posterior variance $\mathrm{Var}_{\mathrm{PA}}[s(\theta, \tilde{Z}) \mid X, Z]$.

Theorem 3.1 works for any function $s(\theta, \tilde{Z})$, but it is useful to know that the same result applies when computing formal posterior distributions of parameters of interest, such as the regression coefficient β_1 in (6). The corollary below extends Theorem 3.1 to this case, as well as any other case where β is some parameter (or set of parameters) of interest.

Corollary 3.1. *Let β be a parameter in the model $p_{\mathrm{SA}}(\theta|\tilde{Z})$ and let $s(\theta, \tilde{Z}) = p_{\mathrm{SA}}(\beta|\theta, \tilde{Z})$. If $\tilde{Z} \subseteq Z$ and $\beta \perp\!\!\!\perp X \mid \theta, \tilde{Z}$, then $s(X, \tilde{Z}, Z)$ is an unbiased estimate of $p_{\mathrm{SA}}(\beta|X, \tilde{Z})$.*

The condition $\beta \perp\!\!\!\perp X \mid \theta, \tilde{Z}$ is essentially guaranteed by design of the measurement process leading to X.

Proof. Observe that

$$\begin{aligned} s(X, \tilde{Z}) &= \int s(\theta, \tilde{Z}) p_{\mathrm{SA}}(\theta|X, \tilde{Z}) \mathrm{d}\theta \\ &= \int p_{\mathrm{SA}}(\beta|\theta, \tilde{Z}) p_{\mathrm{SA}}(\theta|X, \tilde{Z}) \mathrm{d}\theta \\ &= \int p_{\mathrm{SA}}(\beta|\theta, X, \tilde{Z}) p_{\mathrm{SA}}(\theta|X, \tilde{Z}) \mathrm{d}\theta \\ &= p_{\mathrm{SA}}(\beta|X, \tilde{Z}), \end{aligned}$$

where the second to last line follows from the assumption that $\beta \perp\!\!\!\perp X \mid \theta, \tilde{Z}$. □

Theorem 3.1 gives a positive result, in the case that the secondary analyst's $\tilde{Z}$ is a function of the survey institution's Z. Since $\tilde{Z}$ is "invented" by the secondary analyst, however, there is a good chance that $\tilde{Z} \not\subseteq Z$. In that case, the amount of bias is simply

$$E_{\mathrm{SA}}\{E_{\mathrm{PA}}[s(\theta, \tilde{Z})|X, Z]|X, \tilde{Z}\} - E_{\mathrm{SA}}[s(\theta, \tilde{Z})|X, \tilde{Z}]. \tag{11}$$

We generally expect that this bias should decrease as the number of items J in X increases, or equivalently, as the reliability with which θ can be measured by X increases. Mislevy (1991) shows this in the case of a classical true score theory model, and Mislevy et al. (1992) illustrate the same point numerically with an application to SAT testing data. Here, we give an informal argument that this should be true even more generally. Note that the term on the right in (11) is

$$\begin{aligned} E_{\mathrm{SA}}[s(\theta, \tilde{Z})|X, \tilde{Z}] &= \int s(\theta, \tilde{Z}) p(\theta|X, \tilde{Z}) \mathrm{d}\theta \\ &= \int s(\theta, \tilde{Z}) \frac{p(\tilde{Z}|\theta, X)}{p(\tilde{Z}|X)} p(\theta|X) \mathrm{d}\theta \ . \end{aligned}$$

In any measurement model for which there is a consistent estimator $\hat{\theta}(X)$ based on the response variables X, we expect that $\theta \subseteq X$ will become true as J grows (Ellis & Junker, 1997, show a somewhat stronger result for general class of locally independent monotone latent variable models, for example); hence $p(\tilde{Z}|\theta, X)/p(\tilde{Z}|X) \to 1$. Moreover, as J grows, $p(\theta|X)$ should converge to a point mass at the participants' true θ values, θ_{TRUE}.[2] Thus, as $J \to \infty$,

$$\begin{aligned} E_{\mathrm{SA}}[s(\theta, \tilde{Z})|X, \tilde{Z}] &\approx \int s(\theta, \tilde{Z}) p(\theta|X) \mathrm{d}\theta \\ &\to s(\theta_{\mathrm{TRUE}}, \tilde{Z}). \end{aligned}$$

[2] Chang and Stout (1993) give a result implying this for IRT models, and a similar result can be obtained for other psychometric models, by further generalizing the work of Walker (1969).

A similar line of reasoning, beginning with the inner expected value $E_{\mathrm{PA}}[s(\theta, \tilde{Z})|X, Z]$ in the term on the left in (11), shows that this term too converges to $s(\theta_{\mathrm{TRUE}}, \tilde{Z})$ as $J \to \infty$, and hence the bias (11) vanishes as J grows.

Since $\tilde{Z}$ is determined by the secondary analyst long after the survey institution has done the primary analyses, survey institutions try to make Z as large as possible, to accommodate any possible $\tilde{Z}$ that secondary analysts may be interested in. A typical conditioning model (e.g., Kirsch et al., 2000; Mislevy et al., 1992; Dresher, 2006) will involve Z containing all of the variables listed in Sect. 2.2 as well as their two-way interactions. This generally produces a Z with many hundreds of columns. This is reduced by principal components analysis (PCA) to a Z with far fewer columns (e.g., a few hundred), and this is used for all subsequent primary analyses, including the generation of plausible values. Such a large Z is thought to contain a good proxy for any $\tilde{Z}$ that a secondary analyst could define, so that $\tilde{Z} \subseteq Z$ nearly holds, and the bias (11) in $s(X, \tilde{Z}, Z)$ is minimal, even when θ is not measured with high reliability.

Although the construction of such a large Z may seem awkward, it represents an elegant solution to the problem of making primary and secondary analyses logically and arithmetically consistent. For both primary and secondary analysts, computation is simply a matter of summing over plausible values to approximate the functional in (4). If the primary and secondary analysts are using the same set of plausible values, based on a Z designed to contain good proxies for any possible $\tilde{Z}$, then the primary reports are margins of the table of all possible secondary analysis results. If we are able to aggregate across secondary analyses to produce a reporting quantity such as the mean proficiency for female students, this must produce the same answer as the primary analysis did by calculating that mean directly, since it amounts to summing across plausible values in a different order. Thus, any inconsistencies between primary and secondary analyses must be due to arithmetic errors, or conceptual errors in setting up the quantity to be computed, rather than differences in computational methods or tools.

Finally, we note in passing that making Z much larger than $\tilde{Z}$ causes some inefficiency, as can be seen by comparing the variability of $s(X, \tilde{Z}, Z)$ with that of $s(X, \tilde{Z})$ over random replications of the survey,

$$\begin{aligned}\mathrm{Var}\left(s(X, \tilde{Z}, Z)\right) &= E\left[\mathrm{Var}\left(s(X, \tilde{Z}, Z)\middle| X, \tilde{Z}\right)\right] + \mathrm{Var}\left(E\left[s(X, \tilde{Z}, Z)\middle| X, \tilde{Z}\right]\right) \\ &= E\left[\mathrm{Var}\left(s(X, \tilde{Z}, Z)\middle| X, \tilde{Z}\right)\right] + \mathrm{Var}\left(s(X, \tilde{Z})\right),\end{aligned}$$

where the last line follows directly if $\tilde{Z} \subseteq Z$. However, since there are no replications of surveys in practice, this inefficiency is usually overlooked.

4. The θ-Independent Case

Suppose now that the secondary analyst has a research model of the form of (7), namely

$$p_{\mathrm{SA}}(Y|\theta, \tilde{Z}),$$

in which θ now plays the role of an independent variable, and again the secondary analyst is interested in a quantity of the form $s(\theta, Y, \tilde{Z})$. Once again, θ is completely missing, and so it is natural to consider a marginal quantity like

$$s(X, Y, \tilde{Z}) = \int s(\theta, Y, \tilde{Z})\, p_{\mathrm{SA}}(\theta|X, Y, \tilde{Z})\mathrm{d}\theta.$$

By replacing $\tilde{Z}$ with $(Y, \tilde{Z})$, we immediately obtain natural corollaries to Theorem 3.1 and Corollary 3.1. For these corollaries, stated below, we also define

$$s(X, Y, \tilde{Z}, Z) = \int s(\theta, Y, \tilde{Z}) p_{\mathrm{PA}}(\theta|X, Z)\mathrm{d}\theta$$

for the institutional posterior distribution $p_{\mathrm{PA}}(\theta|X, Z)$, perhaps available to secondary analysts through the publication of plausible values. We then immediately obtain

Corollary 4.1. *If $(Y, \tilde{Z}) \subseteq Z$, then $s(X, Y, \tilde{Z}, Z)$ is an unbiased estimate of $s(X, Y, \tilde{Z})$.*

Corollary 4.2. *Let β be a parameter in the model $p_{\mathrm{SA}}(Y|\theta, \tilde{Z})$ and let $s(\theta, Y, \tilde{Z}) = p(\beta|\theta, Y, \tilde{Z})$. If $(Y, \tilde{Z}) \subseteq Z$ and $\beta \perp\!\!\!\perp X \mid \theta, Y, \tilde{Z}$, then $s(X, Y, \tilde{Z}, Z)$ is an unbiased estimate of $p(\beta|X, Y, \tilde{Z})$.*

Corollaries 4.1 and 4.2 assert that Y, which is already a dependent variable in the secondary analyst's model $p_{\mathrm{SA}}(Y|\theta, \tilde{Z})$, should also be an independent variable in the primary analyst's conditioning model $p_{\mathrm{PA}}(\theta|Z)$, *in order that $S(X, Y, \tilde{Z}, Z)$—or its approximation using institutional PVs—produces an unbiased estimate* of $s(X, Y, \tilde{Z})$. However, the assumption that Y is in the institutional conditioning model imposes a serious restriction on what the secondary analyst's model can be; violating this constraint can produce other biases.

Indeed Theorem 4.1, which we present next, shows that when Y is in the institutional conditioning model, the form of $p(Y|\theta, \tilde{Z})$ is essentially determined by the institutional conditioning model. If the form of the secondary analyst's research model $p_{\mathrm{SA}}(Y|\theta, \tilde{Z})$ is the same as the form determined by Theorem 4.1 from the institutional conditioning model $p(\theta|Y, \tilde{Z})$, then $s(X, Y, \tilde{Z}, Z)$—and, therefore, its estimate using institutional PVs—will be an unbiased estimate of $s(X, Y, \tilde{Z})$. If not, the PV-based approximation to $s(X, Y, \tilde{Z}, Z)$ will be vulnerable to bias, as an estimate of $s(X, Y, \tilde{Z})$.

We illustrate this bias in Sect. 6 below. There, we suppose that the secondary analyst's research model $p_{\mathrm{SA}}(Y|\theta, \tilde{Z})$ is in the form of the *wage equation* from economics, which in our case is a linear regression model for $Y = \log(\text{wage})$ in terms of reading proficiency θ and additional covariates $\tilde{Z}$ that include race/ethnicity and work experience. We attempt to estimate regression coefficients in the wage equation using PVs released for the NALS. In this case, the primary NALS analysis explicitly includes (a function of) Y in the institutional conditioning model, but the form for $p_{\mathrm{SA}}(Y|\theta, \tilde{Z})$ determined from the conditioning model by Theorem 4.1 is undoubtedly different from the wage equation in the economics literature. The estimates of coefficients in the wage equation based on NALS PVs are thus vulnerable to bias. We examine the size and direction of the bias, in Sect. 6, by comparing PV-based coefficient estimates with estimates that do not depend on institutional PVs.

For ease of exposition, we consider in Theorem 4.1 the case in which $Z = (Y, \tilde{Z})$. However, as discussed below with respect to Corollary 4.3(b) and illustrated in Sect. 6, we can expect similar behavior in the more general case $Z \supseteq (Y, \tilde{Z})$. Note that in Theorem 4.1, $p(\theta|Y, \tilde{Z})$ is the institutional conditioning model as in (2), not a posterior distribution, and $p(Y|\theta, \tilde{Z})$ is the form that the secondary analyst's research model $p_{SA}(Y|\theta, \tilde{Z})$ must match.

Theorem 4.1. *The distribution $p(Y|\theta, \tilde{Z})$ is completely determined by the conditioning model $p(\theta|Y, \tilde{Z})$ and the conditional distribution $p(Y|\tilde{Z})$.*

Since $p(Y|\tilde{Z})$ is entirely determined by the observable relationship between Y and $\tilde{Z}$, $p(Y|\theta, \tilde{Z})$ is essentially determined once $p(\theta|Y, \tilde{Z})$ is specified.

Proof. Observe that

$$p(Y|\theta, \tilde{Z}) = \frac{p(\theta|Y, \tilde{Z})}{p(\theta|\tilde{Z})} \cdot p(Y|\tilde{Z}). \tag{12}$$

For the denominator in (12), we note

$$p(\theta|\tilde{Z}) = \int p(\theta, Y|\tilde{Z})\mathrm{d}Y = \int p(\theta|Y, \tilde{Z})p(Y|\tilde{Z})\mathrm{d}Y. \tag{13}$$

Clearly, Eqs. (12) and (13) depend only on $p(\theta|Y, \tilde{Z})$ and $p(Y|\tilde{Z})$, and completely determine $p(Y|\theta, \tilde{Z})$. □

Now let us consider a more general conditioning model $p(\theta|U, \tilde{Z})$. In the special case of Theorem 4.1, $U = Y$. Or, U might be a variable specifically intended to proxy for Y. Or, if we have a large institutional conditioning model $p(\theta|Z)$ with $\tilde{Z} \subseteq Z$, U might be all the information in Z that is not in $\tilde{Z}$, in which case U might be highly multivariate.

Part (a) of Corollary 4.3, which we present next, shows that if the relationship between U and Y is completely explained by θ and $\tilde{Z}$—in the sense that $Y \perp\!\!\!\perp U|\theta, \tilde{Z}$—then adding U to the conditioning model does not force a particular form for the secondary analyst's research model $p(Y|\theta, \tilde{Z})$. For example, in the wage equation example of Sect. 6, an indicator variable U that is 1 for someone who has been working 5–10 years and 0 otherwise is undoubtedly associated with $Y = \log(\text{wage})$, but since U is a function of the work experience variable that is already in $\tilde{Z}$, including U in the conditioning model does not constrain the specification of the wage equation by the secondary analyst.

On the other hand, part (b) of Corollary 4.3 shows that if $Y \not\perp\!\!\!\perp U|\theta, \tilde{Z}$, then once again the form of $p(Y|\theta, \tilde{Z})$ is determined by the conditioning model $p(\theta|U, \tilde{Z})$. For example, if we have a large institutional conditioning model that is designed so that Z contains "proxies for everything," and U contains all of the information in Z that is not in $\tilde{Z}$, then it is very likely that $Y \not\perp\!\!\!\perp U|\theta, \tilde{Z}$. Hence, the secondary analyst's research model $p_{\mathrm{PA}}(Y|\theta, \tilde{Z})$ must match the form determined by the conditioning model $p(\theta|Z)$.

Corollary 4.3.

(a) If $Y \perp\!\!\!\perp U|\theta, \tilde{Z}$, then the conditioning model $p(\theta|U, \tilde{Z})$ places no constraint on the distribution $p(Y|\theta, \tilde{Z})$.

(b) If $Y \not\perp\!\!\!\perp U|\theta, \tilde{Z}$, then the conditioning model $p(\theta|U, \tilde{Z})$ and the conditional distribution $p(U|\tilde{Z})$ determine the distribution $p(Y|\theta, \tilde{Z})$.

Proof. We first observe that, as in the proof of Theorem 4.1,

$$p(U|\theta, \tilde{Z}) = \frac{p(\theta|U, \tilde{Z})}{p(\theta|\tilde{Z})} \cdot p(U|\tilde{Z}),$$

which again depends only on $p(\theta|U, \tilde{Z})$ and $p(U|Z)$. Then,

$$\begin{aligned} p(Y|\theta, \tilde{Z}) &= \int p(Y, U|\theta, \tilde{Z})\mathrm{d}U \\ &= \int p(Y|\theta, U, \tilde{Z})p(U|\theta, \tilde{Z})\mathrm{d}U. \end{aligned} \tag{14}$$

If $Y \perp\!\!\!\perp U|\theta, \tilde{Z}$ then the first term under the integral in (14) reduces to $p(Y|\theta, \tilde{Z})$, and there is no constraint. However, if $Y \not\perp\!\!\!\perp U|\theta, \tilde{Z}$, then (14) determines $p(Y|\theta, \tilde{Z})$. □

Corollaries 4.1 and 4.2 show that, in order to use institutional PV methodology to explore predictive inference using θ and other covariates, the variable Y to be predicted—or a good proxy of it—must be in the institutional conditioning model. But Theorem 4.1 and Corollary 4.3 show that to include Y or a non-trivial proxy U forces a particular form for the secondary analyst's research model. When this form is not used, bias may result, as illustrated below in Sect. 6, and this may lead to incorrect scientific or policy conclusions. It might be reassuring to realize that the bias should decrease as test length increases (or measurement error for θ decreases), as suggested by the discussion of Eq. (11), but the direction and magnitude of the bias will vary depending on the application.

Since survey institutions typically release only the plausible values and not the details of the model associated with them, it is unlikely that the secondary analyst could specify $p_{SA}(Y|\theta, \tilde{Z})$ to match the form suggested by the conditioning model, even if he/she wished to deviate from the research models in the substantive literature (such as the wage equation in economics). Thus, for predictive inference using θ, the secondary analyst is better off building a model from scratch, not making use of institutional PVs. We turn to this process in the next section.

5. A Workable Approach to the θ-Independent Case

In Sect. 4, we argued that the usual plausible values methodology, using institutional PVs generated from a large, fixed conditioning model in order to calculate unbiased estimates of $s(X, Y, \tilde{Z})$, is not usually a tenable practice. An alternative would be to build a model directly for the secondary analyst's research question. The following easy proposition summarizes the essential features of a marginal likelihood or Bayesian model built by the secondary analyst.

Proposition 5.1. *Let β be any parameter(s) of interest. Then, under the setup of Sect. 2, if $X \perp\!\!\!\perp \tilde{Z}, \beta|\theta$ and $\theta \perp\!\!\!\perp \beta|\tilde{Z}$,*

(a) The secondary analyst's marginal likelihood for β is

$$p_{\mathrm{SA}}(X, Y|\tilde{Z}, \beta) = \int p_{\mathrm{SA}}(Y|\theta, \tilde{Z}, \beta)p_{\mathrm{SA}}(X|\theta)p_{\mathrm{SA}}(\theta|\tilde{Z})\mathrm{d}\theta.$$

(b) The secondary analyst's posterior distribution for β is

$$p_{\mathrm{SA}}(\beta|X, Y, \tilde{Z}) = \frac{\int p_{\mathrm{SA}}(Y|\theta, \tilde{Z}, \beta)p_{\mathrm{SA}}(X|\theta)p_{\mathrm{SA}}(\theta|\tilde{Z})\mathrm{d}\theta p_{\mathrm{SA}}(\beta)}{\int\int p_{\mathrm{SA}}(Y|\theta, \tilde{Z}, \beta)p_{\mathrm{SA}}(X|\theta)p_{\mathrm{SA}}(\theta|\tilde{Z})\mathrm{d}\theta p_{\mathrm{SA}}(\beta)\mathrm{d}\beta}.$$

The condition $X \perp\!\!\!\perp \tilde{Z}, \beta|\theta$ is essentially guaranteed by good measurement practice in constructing X as a measure of θ. The condition $\theta \perp\!\!\!\perp \beta|\tilde{Z}$ is quite innocuous in the θ-independent case, as we shall see after the proof.

Proof. For part (a), we observe

$$\begin{aligned} p_{\mathrm{SA}}(X, Y|\tilde{Z}, \beta) &= \int p_{\mathrm{SA}}(X, Y, \theta|\tilde{Z}, \beta)\mathrm{d}\theta \\ &= \int p_{\mathrm{SA}}(Y|\theta, \tilde{Z}, \beta)p_{\mathrm{SA}}(X|\theta, \tilde{Z}, \beta)p_{\mathrm{SA}}(\theta|\tilde{Z}, \beta)\mathrm{d}\theta \\ &= \int p_{\mathrm{SA}}(Y|\theta, \tilde{Z}, \beta)p_{\mathrm{SA}}(X|\theta)p_{\mathrm{SA}}(\theta|\tilde{Z})\mathrm{d}\theta, \end{aligned}$$

where the second line follows from the law of total probability and the third line follows from the conditional independence assumptions.

For part (b) we observe that

$$p_{\mathrm{SA}}(\beta|X,Y,\tilde{Z}) = \frac{p_{\mathrm{SA}}(\beta,X,Y|\tilde{Z})}{p_{\mathrm{SA}}(X,Y|\tilde{Z})} = \frac{p_{\mathrm{SA}}(X,Y|\tilde{Z},\beta)p_{\mathrm{SA}}(\beta)}{\int p_{\mathrm{SA}}(X,Y|\tilde{Z},\beta)p_{\mathrm{SA}}(\beta)\mathrm{d}\beta}.$$

The result now follows by applying part (a). □

Note that Y is forced to be in the institutional PV conditioning model in Corollary 4.1, but it does not appear in the secondary analyst's "conditioning model" $p(\theta|\tilde{Z})$ in Proposition 5.1. There is a subtle but important distinction between the problems that these two results solve. Corollary 4.1 shows what is necessary for institutional PVs so that $s(X,Y,\tilde{Z},Z)$ (or its PV-based approximation) is an unbiased estimate of $s(X,Y,\tilde{Z})$. Proposition 5.1 shows what is necessary for the rather different problem of making inferences about β directly from the data and tools available to the secondary analyst, *ignoring* the institutional model used for primary analyses. It is still necessary for the secondary analyst to incorporate appropriate survey design variables into $\tilde{Z}$, in order to avoid omitted-variable biases for example, but Y plays a rather different role in the model of Proposition 5.1, versus the institutional PV conditioning model.

The model implied by Proposition 5.1 can be instantiated, for example, as a mixed effects structural equations (MESE) model (Schofield, 2008; Junker, Schofield, & Taylor, 2012). The MESE model takes the form

$$Y_i|\tilde{Z}_i,\theta_i,\beta,\sigma \sim p(Y_i|\theta_i,\tilde{Z}_i,\beta,\sigma), \tag{15}$$

$$X_{ij}|\theta_i,\gamma_j \sim p(X_{ij}|\theta_i,\gamma_j), \tag{16}$$

$$\theta_i|\tilde{Z}_i,\alpha,\tau \sim p(\theta_i|\tilde{Z}_i,\alpha\ ,\tau), \tag{17}$$

where θ_i, $\tilde{Z}_i$, and Y_i are defined as before.

Although it is expressed in hierarchical Bayes form here, it is not difficult to see that the MESE model is a generalized structural equations model (SEM; e.g., Bollen, 2002). In particular, it is an extension of the MIMIC model (Joreskog & Goldberger, 1975) that allows for general data types and additional observed covariates. The MESE model allows for examination of the conditional effect of $\tilde{Z}$ on the dependent variable Y after controlling for θ, through the parameter(s) β. It is a "mixed effects" model, because the latent variable is modeled as a random effect, while the other covariates in (15) are fixed effects.

Equation (15) corresponds to the secondary analyst's model and is of primary substantive interest. Often (15) will take the form of a generalized linear model. For example, Junker et al. (2012) apply MESE models with linear regression components such as

$$Y_i = \beta_0 + \beta_1\theta_i + \beta_2\tilde{Z}_i + \varepsilon_i, \quad \varepsilon_i \sim N(0,\sigma^2), \tag{18}$$

and with logistic regression components, in (15).

Equation (16) is the measurement model for θ with parameters γ_j. Frequently, (16) will be an IRT or an MIRT model, but any measurement model (e.g., a classical test theory model, a factor analysis model, a cognitive diagnosis model) is possible. When the survey institution has built and calibrated a reliable measurement model, for example an IRT model, then it is best for the secondary analyst to use that model and those calibration estimates of γ_j in (16).

Equation (17) is a prior for θ; it is the secondary analyst's "conditioning model." But, as argued above, because the problem being solved now is different from the one that institutional PVs were designed to solve, (17) must have a form different from the institutional conditioning model. In the MESE model, the conditioning model plays the role of a prior on θ and it must condition on $\tilde{Z}$ that appear as covariates in (15), but does not condition on dependent variable Y.

Note that (16) uses item response data, which the survey institution must make available to secondary researchers. While these responses are available for some surveys (NCES provides item responses for those researchers who have restricted use data licenses in many surveys), many other surveys (e.g., the 1979 and 1997 National Longitudinal Survey of Youth) do not publish item responses for all cognitive tests. *Items* need not be released, but rather *item responses* along with information about the construction and design of the test (e.g., item parameters and information on the model used to construct the test) are sufficient for estimation and inference with the MESE model.

6. Data Example: Racial Wage Gaps in the 1992 National Adult Literacy Survey

The wage regression literature in econometrics provides an important example of the analysis we have been discussing. A typical problem in this area examines wage gap between two social groups; the typical solution is to build a regression equation that accounts for different factors that could affect the wage differential, while controlling for cognitive proficiency through test scores.

To illustrate our theoretical results, we examine the issue of racial wage gaps with the hope of decomposing the gaps in order to understand the impact of the racial differences in cognitive proficiency on differences in labor market earnings. The model of interest is

$$log(\text{wage}_i) = \beta_0 + \beta_1\theta_i + \beta_2\text{black}_i + \beta_3\text{Hispanic}_i + \beta_4 W_i + \varepsilon_i, \tag{19}$$

where wage_i is weekly wage rate for individual i, θ_i is a measure of cognitive proficiency (in this case English-language prose literacy), black_i and Hispanic_i are indicator variables for race/ethnicity identification (with non-Hispanic white serving as the omitted category), and W is a vector that includes potential labor market experience, region, and urban status. Here, the variables black_i, Hispanic_i, and W_i comprise $\tilde{Z}$.

Equation (19) is a wage regression of the form seen in the literature (e.g., Neal & Johnson, 1996; Lang & Manove, 2011; Junker et al., 2012).[3] Of primary interest are estimates of the regression coefficients in front of the racial indicator variables ($\hat{\beta}_2$ and $\hat{\beta}_3$) as they provide information on racial differences in labor market earnings after controlling for differences in cognitive proficiency. Accurate estimates can potentially inform policy, with the goal of targeting appropriate resources to address the earnings differential.

The data come from the 1992 (NALS; Kirsch et al., 2000) which includes an individually administered household survey of 24,944 adults ages 16 and over. The NALS is comprised of two sets of questions—standard demographic questions (e.g., race, gender, labor force behavior, wages, age, etc.) and items that measure functional literacy in three unidimensional domains: prose, document, and quantitative. For completeness, we describe the design in terms of all three domains, but in our analyses below we include only a latent variable θ for the unidimensional prose domain.

[3] The basic form of Eq. (19) is very common in the labor economics literature. However, the work we present below differs in many ways from Neal and Johnson (1996) and Lang and Manove (2011). Those papers use different data than we do, they use a different measure of cognitive ability (and that measure is constructed when individuals are teenagers, while wages are measured many years later when individuals are working adults), and their regressions include different covariates than ours. Thus, while the basic form of the regression is standard, interpretation of results will differ.

TABLE 1.
Sample characteristics of the 1992 NALS data.

	Black	Hispanic	White
N	622	470	3858
Average age	39.5	37.5	40.5
Average weekly wage	479.03	468.41	721.25
Average prose literacy scores	−0.36	−0.62	0.75

Authors' calculations, 1992 NALS. Data are from the 1992 NALS, restricted to individuals aged 25–65 who work full time, reported wages, and answered at least one prose literacy item.

The NALS contains 165 items (41 prose literacy, 81 document literacy, and 43 quantitative literacy) to test the literacy skills of the examinees. Because it was deemed impractical to administer every item to every respondent, each respondent was randomly administered a booklet designed via the balanced incomplete block (BIB; Beaton & Zwick, 1992) spiral design which contained a representative sub-sample of approximately one-quarter to one-third of the full set of 165 items. On the prose scale, each individual received a booklet that contained between eight to thirteen prose items rather than the full 41 items.

Since the primary purpose for conducting the NALS survey is to provide information on the literacy skills of US adults, population estimates are of more interest than individual-level literacy scores. Consequently, the NALS dataset does not contain individual literacy proficiency estimates, but instead contains five PVs per content area and individual to aid in calculating population estimates through the marginal estimation procedures outlined in Sect. 3 (with further information available in Mislevy 1991).

Kirsch et al. (2000) describe the conditioning model used in the production of plausible values for the NALS data. The NALS conditioning model is "a normal multivariate distribution …with a common variance, S, and with a mean given by a linear model with slope parameters, G, based on the first approximately 100 principal components of several hundred selected main effects and two-way interactions of the complete vector of background variables" (Kirsch et al., 2000, p. 180). The full list of background variables placed in the principal component analysis is available in Appendix B of the report. For our purposes, it is important to note this saturated model used in producing the plausible values contains wages—the dependent variable in our analysis—and several other variables not in (19), the regression of interest. Many of these additional variables are correlated with wages such as family income and occupation.

We compare estimates of the coefficients on prose literacy (β_1), black (β_2), and Hispanic (β_3) in (19) under several different model specifications, using a restricted sample consisting of the 4950 men in the study who are between the ages of 25 and 65, work full time, report a wage, and who self-report as black, Hispanic, or non-Hispanic white. Sample characteristics are available in Table 1. We note a few features of the data. On average, white men earn more than black and Hispanic men and white men have higher average prose literacy scores than black men or Hispanic men.

We report estimates of the regression coefficients for three different model specifications in Table 2. In the first specification, the explanatory variables include two indicator variables for racial identification (black and Hispanic), potential experience entered as a quadratic where potential experience is defined as the approximate number of possible years in the labor force (defined *age—years of schooling—6*), and indicator variables for census regions (midwest, south, and west) and urban settings. Column (a), provides a "baseline" for the black–white and Hispanic white wage gaps without controlling for prose literacy proficiency. We expect to see

omitted-variable bias in the regression coefficients in column (a), because an important variable—prose literacy—has been omitted from the model.

The next two specifications control for prose literacy proficiency but do so in different ways. In column (b), we include a measure of prose literacy proficiency using the plausible values provided by the survey institution with the saturated conditioning model. We follow Mislevy (1991) and Kirsch et al. (2000) by estimating the regression five times (each with a different prose plausible value as the measure of θ). The best estimate of the regression coefficients is the average of the five regression coefficients obtained from the analyses using different sets of plausible values. We estimate the standard errors of the regression coefficients by using the between/within calculation recommended by Mislevy (1991, p. 182); see also the discussion of this calculation in Sect. 3 above.

Since Y is in the conditioning model used to generate the institutional PVs, Theorem 4.1 and Corollary 4.3 suggest that the conditioning model will determine a unique form for $p(Y|X, \theta, \text{black}, \text{Hisp}, W)$. Because of the complex nature of the conditioning model used by the survey institution, it is impossible for us to determine the functional form of $p(Y|X, \theta, \text{black}, \text{Hisp}, W)$ determined by the NALS conditioning model. Additionally, it is highly unlikely that our wage regression shown in (19) will be compatible with the survey institution's $p(Y|X, \theta, \text{black}, \text{Hisp}, W)$. For these reasons, we expect to see bias in our estimates of β if we use the survey institution's PV's in our secondary analysis.

In the MESE specification, shown in column (c), we estimate the regression coefficients using the (MESE) model described in Sect. 5. We report estimates from a MESE model where we specify the conditioning model to include only race, experience, census region, and urban setting. For this example, the MESE model takes the form

$$Y_i|R_i, W_i, \theta_i \sim N(\beta_0 + \beta_1\theta_i + \beta_2\text{black}_i + \beta_3\text{Hispanic}_i + \beta_4 W_i, \sigma^2), \tag{20}$$

$$X_{ij}|\theta_i \sim IRT(X_{ij}|\theta_i, \gamma_j,) \tag{21}$$

$$\theta_i|R_i, W_i \sim N(\theta_i|\alpha_0 + \alpha_1\text{black}_i + \alpha_2\text{Hispanic}_i + \alpha_3 W_i, \tau^2), \tag{22}$$

where y_i is log(wages), X_{ij} are prose literacy item responses, θ is latent prose literacy proficiency, R_i are the race/ethnicity variables, and W_i are the additional variables we control for that include experience, census region and urban status indicator variables. [R_i and W_i are the components of $\tilde{Z}_i$ in Eqs. (15) and (17).]

In the MESE specification, we use the same 3-PL IRT model specified by Kirsch et al. (2000) as the measurement model for prose literacy in the primary analyses, and we fix the item parameters at the estimates reported in Kirsch et al. (2000, Appendix A), to ensure that the differences in the MESE versus PV estimates are not due to measurement model differences.

Kirsch et al. (2000) note that the IRT model used in the NALS primary analyses is a multiunidimensional model that treats each scale independently and so "a unique proficiency was assumed for each scale" (p. 170) and all items load on only scale. We include only those items deemed "prose" items in the MESE model, and compare our results to results using institutional PVs for prose literacy proficiency only. While the conditioning model used in the production of the PVs is multivariate, separate estimates of the conditioning variables were estimated for each scale, so the additional items on the document and quantitative literacy scale effectively function as additional conditioning variables in the conditioning model for prose literacy. Because the MESE model is correctly specified as described in Sect. 5 and because it does not depend on institutional plausible values, we do not expect it to exhibit the biases that the PV-based methodology exhibits.

We use a Markov Chain Monte Carlo (MCMC) procedure to estimate the regression coefficients in the MESE specification using WinBUGS software (Spiegelhalter, Thomas, & Best, 2000). R and WinBUGS code implementing this MESE model is available from the authors.

TABLE 2.
Wage regressions comparing NCES PVs and MESE using 1992 NALS data.

	No θ	PV	MESE
Conditioning Model contains Y?		Yes	No
	(a)	(b)	(c)
Black	−0.368	−0.140	−0.083
	(0.027)	(0.027)	(0.029)
Hispanic	−0.453	−0.145	−0.172
	(0.030)	(0.031)	(0.033)
92 NALS prose lit		0.247	0.256
		(0.010)	(0.013)
N	4,950	4,950	4,950

All regressions control for potential experience entered as a quartic, census region (entered as dummy variables), and urban setting (entered as a dummy variable). PV estimates in column (b) employ the recommended procedure (Mislevy, 1991) for combining regression results for multiple imputations. PV estimates use the NCES saturated conditioning set which includes wage (the dependent variable) variables correlated with wages, race, experience, census region, and urban setting. The MESE model estimates in column (c) are posterior means reported from estimating a MESE model. In column (c), the conditioning set includes race, experience, census region, and urban setting. Literacy has been scaled such that it is the same in columns (b)–(c) for comparison purposes.

Table 2 shows substantially different results depending on the conditioning model used. Based on arguments made in Sects. 4 and 5, we have reason to suspect that there are biases in the estimates of the regression coefficients in column (b), because the model in our secondary analysis is not compatible with the primary institution's conditioning model.

The coefficient on the variable black is estimated to be substantially lower using the MESE model than with the PV methodology. The MESE estimates indicate that prose literacy proficiency "accounts for" approximately 77 % (1-0.083/0.368) of the black-white log wage gap. By way of comparison the results in the regression estimated by PV methodology (which is likely biased) suggests that literacy proficiency accounts for only 62 % (1-0.140/0.368) of this gap.

Also, using the PV-based model, we would infer that Hispanics and blacks have a similar wage disadvantage relative to whites, once we account for prose literacy proficiency. In contrast, the MESE estimates suggests that among men with comparable literacy proficiency, the wage gap is higher for Hispanic men than black men.

Our example is intended as an exercise to demonstrate differences in inference that can arise from differences in methodological approach. We note, though, that our example comes from a literature that has an important goal—providing useful policy guidance. For instance, if empirical work establishes that racial and ethnic wage gaps are primarily the consequence of differences in proficiencies (literacy, as in our example, or other skills), this suggests that there is an important role for allocating educational resources so as to narrow racial and ethnic differences in the acquisition of skills. If, instead, wage gaps are found to persist even among similarly skilled individuals, this might direct a focus to policies designed to reduce discrimination in labor markets.

This empirical exercise demonstrates the biases that can result from using institutional PVs in a secondary analysis that is not compatible with the conditioning model used to procure the PVs. Our secondary analysis is not a rigged example; regressions of the form (19) are used in many studies in labor economics (as well as in other areas of the social sciences). Using real data, we have demonstrated what we argued in Sect. 4 analytically: institutional PVs will produce biased estimates of regression coefficients when the secondary analysis is not compatible with the survey institution's conditioning model.

7. Discussion

Institutional plausible values (PVs) are multiple imputations from an institutional posterior distribution $p(\theta|X, Z)$ based on a large latent regression model known as a "conditioning model." Institutional PVs were designed for, and are successful at, allowing secondary analysts to produce unbiased estimates of quantities associated with a "small" posterior distribution $p(\theta|X, \tilde{Z})$ related to the secondary analyst's research questions, using machinery from the larger institutional posterior distribution. This situation, which we call the θ-*dependent* case, covers all primary and much secondary analysis of education survey data: characterizing θ in one or more subgroups of interest. In the first part of this paper, we have reformulated basic results for this case, going back to Mislevy (1991), in modern and fairly general notation. We have shown that the utility of PVs is likely to extend across a much larger range of measurement and conditioning models than has been considered previously in studies of PV methodology.

We have also considered the θ-*independent* case, in which an outcome variable Y is predicted from θ and other covariates $\tilde{Z}$. Such models are often expressed as regressions for Y on $\tilde{Z}$ and θ, and can be found in many social science settings, from predicting end-of-year exams using progress in a tutoring system throughout the year (Ayers & Junker, 2008) to understanding social group-based gaps in wages after accounting for cognitive status (Junker et al., 2012), to predicting post-secondary major choice and success (Schofield, 2013).

Unfortunately, our results show—both theoretically and empirically—that standard institutional PVs cannot lead to the same sorts of unbiased estimates of posterior quantities in the θ-independent case. Indeed, the only time that unbiased results are guaranteed using standard PV methodology in the θ-independent case is if the institution releases enough information that a secondary analyst could build a θ-independent model compatible with the conditioning model that generated the PVs.

There is, however, a fix at hand. Instead of trying to use PVs in the θ-independent case, secondary analysts can build the relevant model for their research questions, from scratch, largely ignoring the institutional machinery. One such model, the Mixed Effects Structural Equations (MESE) model, offers considerable flexibility, great initial success, and can account carefully for the latent structure and its unreliable measurement.

In order to use the MESE model or any related model, the secondary analyst must have access to the individual item responses of survey participants on the cognitive portion of the survey, and ideally also the measurement model that the survey institution used to build the θ scale. The advantage of doing so is that secondary analyst can take advantage of the measurement model that the survey institution designed to characterize carefully and account for the measurement uncertainty in the θ-independent model. Other approaches—such as classical errors-in-variables or instrumental variables approaches—require additional, often untested and untenable, assumptions to account for the measurement error.

Finally, while we focused our discussion and example on IRT models and linear regression models, there is nothing in our analysis that requires this structure. The points we raise here are issues for a wide class of models and a wide class of latent variables. The proper conditioning set for latent variables used as covariates is necessary for the consistency of estimates.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper. The work was supported by Award Number R21HD069778 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development. The content is solely the responsibility of the authors and does not necessarily

represent the official views of the Eunice Kennedy Shriver National Institute of Child Health and Human Development or the National Institutes of Health.

References

Allen, N. L., Carlson, J. E., & Zelenak, C. A. (1999). *The NAEP 1996 Technical Report (NCES 99452)*. Washington, DC: National Center for Education Statistics.

Allen, N. L., Donoghue, J. R., & Schoeps, T. L. (2001). T*he NAEP 1998 Technical Report (NCES 2001–509)*. Washington, DC: National Center for Education Statistics. Retrieved from http://nces.ed.gov/nationsreportcard/pubs/main1998/2001509.asp.

Ayers, E., & Junker, B. (2008). IRT modeling of tutor performance to predict end-of-year exam scores. *Educational and Psychological Measurement*, *68*, 972–987.

Beaton, A. E., & Zwick, R. (1992). Overview of the National Assessment of Educational Progress. *Journal of Educational Statistics*, *17*(2), 95–109. Special Issue: National Assessment of Educational Progress.

Billingsley, P. (1986). *Probability and measure*. New York, NY: Wiley.

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology*, *55*, 605–634.

Carstens, R. & Hastedt, D. (2010). *The effect of not using plausible values when they should be: An illustration using TIMSS 2007 grade 8 mathematics data*. Paper presented at the 4th IEA International Research Conference (IRC-2010), Gothenburg, Sweden. Retrieved July 2013 from http://www.iea.nl/irc-2010.html.

Chang, H., & Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika*, *58*, 37–52.

Dresher, A. (2006). *Results from NAEP marginal estimation research*. Presented at the annual meeting of the National Council on Measurement in Education (NCME), San Francisco, CA.

Ellis, J. L., & Junker, B. W. (1997). Tail-measurability in monotone latent variable models. *Psychometrika*, *62*, 495523.

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. New York, NY: Springer.

Johnson, M.S., & Jenkins, F. (2005). *A Bayesian hierarchical model for large-scale educational surveys: An application to the National Assessment of Educational Progress*. Research Report #RR-04-38. Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/research/policy_research_reports/publications/report/2005/hyzl.

Joreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*, 631–639.

Junker, B. W., Schofield, L. S., & Taylor, L. (2012). The use of cognitive ability measures as explanatory variables in regression analysis. *IZA Journal of Labor Economics*, *1*(4), 4.

Kirsch, I., et al. (2000). *Technical report and data file user's manual for the 1992 National Adult Literacy Survey*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Lang, K., & Manove, M. (2011). Education and labor market discrimination. *American Economic Review*, *101*, 1467–1496.

Li, D., Oranje, A., & Jiang, Y. (2009). On the estimation of hierarchical latent regression models for large-scale assessments. *Journal of Educational and Behavioral Statistics*, *34*, 433–463.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, *9*(4), 538–558.

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

Mislevy, R. J. (1993). Should "multiple imputations" be treated as "multiple indicators"? *Psychometrika*, *58*, 79–85.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*(2) NAEP, 133–161.

NCES (2008). *NAEP secondary analysis grant abstracts*. Washington, DC: National Center for Education Statistics. Retrieved from http://nces.ed.gov/nationsreportcard/researchcenter/naepgrants.asp.

NCES (2009). *NAEP technical documentation*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics (NCES). Retrieved July 2013 from http://nces.ed.gov/nationsreportcard/tdw/.

Neal, D., & Johnson, W. (1996). The role of pre-market factors in black-white wage differences. *Journal of Political Economy*, *104*, 869–895.

Olson, J. F., Martin, M. O., & Mullis, I. V. S. (Eds.). (2008). *TIMSS 2007 technical report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Robitaille, D. F., & Beaton, A. E. (Eds.). (2002). *Secondary analysis of the TIMSS data*. New York, NY: Springer.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association*, *9*, 473–489.

Schofield, L. S. (2008). *Modeling measurement error when using cognitive test scores in social science research*. Doctoral dissertation. Pittsburgh, PA: Department of Statistics and Heinz College of Public Policy, Carnegie Mellon University.

Schofield, L. S. (2013). Measurement error in latent variables that predict STEM retention (under review).

Spiegelhalter, D. J., Thomas, A., & Best, N. G. (2000). *WinBUGS Version 1.3 User Manual*. Cambridge: Medical Research Council Biostatistics Unit. Retrieved from http://www.mrc-bsu.cam.ac.uk/bugs.

van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York, NY: Springer.

von Davier, M., & Gonzalez, M. (2009). What are plausible values and why are they useful? In M. von Davier & D. Hastedt (Eds.), *IERI monograph series, vol. 2: Issues and methodologies in large-scale assessments* (pp. 9–36). Princeton,

NJ: International Association for the Evaluation of Educational Achievement (IEA) and Educational Testing Service (ETS).

Walker, A. M. (1969). On the asymptotic behavior of posterior distributions. *Journal of the Royal Statistical Society, Series B*, *31*, 80–88.

Manuscript Received: 29 JAN 2013
Published Online Date: 18 SEP 2014