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 for inputs.

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 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. Precision can be accurately reported 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 & 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 ability 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 allows secondary analysts to account for measurement error in subsequent analyses by employing PVs in appropriate ways (e.g., Mislevy, 1991 and 1993, and 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, Oranje & 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 condi-

tioning model used by the institution to generate PVs. In Section 2 and Section 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. When proficiency is an *independent* variable, however, we provide a disquieting result. In Section 4 we show that secondary analysis is susceptible to substantial wrong-model bias when using institutional PVs 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 wrong-model 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.

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 not the dependent 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 Section 5 we discuss workable machinery to do this.

Our results are stated in considerable generality. Nevertheless we show an example in Section 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 show the wrong-model bias resulting from including Y in the conditioning set of the PVs. We demonstrate the bias that occurs when we do not include the covariates in the conditioning set in our model built from scratch.

2 Modeling Components for the Analysis of Education Surveys

Before discussing the key results of Mislevy (1991) and Mislevy et al. (1992) in Section 3 and exploring their extension to models that use proficiency to predict other outcomes in Section 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 analysis 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, Carlson & Zelenak, 1999; Allen, Donoghue & Schoeps, 2001; Kirsch et al., 2000; NCES, 2009; and Olsen, Martin, & Mullis, 2009). 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 Section 2.1 we review the survey institution's *measurement model*, a generative psychometric model that describes the relationships between particpants' proficiency in a particular cognitive domain, and their responses to particular cognitive items in the survey. In Section 2.2 we review the survey institution's *population model*, also known as the *condi-*

tioning model, which describes variation in cognitive proficiency across groups defined by demographic, jurisdictional, and other background covariates. We then briefly discuss in Section 2.3 the kinds of inference made by primary analysts working on behalf of the survey institution. Finally, in Section 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 i^{th} 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 discrimination a_j and difficulty 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;

- 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 that the γ_j's 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*

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

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;
- Other background variables for individual participants, such as education experience, 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 organizations that participate in the survey. More broadly, it is generally unethical to break confidentiality and privacy commitments typically 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 equations (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

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

under the assumption that $X \perp\!\!\!\perp Z \mid \theta$, which is usually true by design of the measurement process producing X (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, Carlson & Zelenak (1999), Allen, Donoghue & Schoeps (2001), Kirsch, et al. (2000), NCES (2009), and Olsen, Martin, & Mullis (2009)—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]

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

for some appropriate function $s(\theta,Z)$—or they can be derived from functionals of $p_{PA}(\theta|X,Z)$.

[1]As suggested in Section 2.1 and Section 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.

The quantities in (3) and (4) may be estimated using Bayesian methods (Johnson & Jenkins, 2005), or marginal maximum likelihood and other methods (Allen, Donoghue & Schoeps, 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, Gonzalez & Mislevy (2009), and the consequences of their misuse in certain contexts was recently discussed by Carstens & 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

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

where the subscript SA is intended as a reminder that this is a model chosen by the secondary analyst. 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

$p_{SA}(\theta|\tilde{Z})$ 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 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 $p_{SA}(\theta|\tilde{Z})$ of interest. For example in the linear regression case, the secondary analyst might be interested in the least-squares estimate of β_1,

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

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 also 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

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

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 appear 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 $p_{SA}(Y|\theta, \tilde{Z})$ of interest. In the linear regression case, the secondary analyst might be interested in the least-squares estimate of some regression coefficient(s) or the posterior distribution of the regression coefficient(s), for example.

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.

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

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

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

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 analysts provide to the secondary analyst, so that (a) an integral of the form of (9) can be calculated or approximated appro-

priately, 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 Section 2.3.

With institutional PVs the secondary analyst can approximate the quantity

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

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

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

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 & 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 $p_{PA}(\theta|X, Z)$, and then averaging. This produces

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

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 $\mathrm{Var}_{PA}[s(\theta, \tilde{Z}) \mid X, Z]$.

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 $p_{SA}(\theta|\tilde{Z})$ and let $s(\theta, \tilde{Z}) = p_{SA}(\beta|\theta, \tilde{Z})$. 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 $p_{SA}(\beta|X, \tilde{Z})$.*

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

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

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

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

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

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

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

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

[2]Chang & 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).

$J \to \infty$,

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

A similar line of reasoning, beginning with the inner expected value $E_{PA}[s(\theta, \tilde{Z})|X, Z]$ in the term on the left in (11), shows that this term too converges to $s(\theta_{TRUE}, \tilde{Z})$ 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 accomodate 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 Section 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 just a few hundred columns 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 from examining the variability of $s(X, \tilde{Z}, Z)$ over random replications of the survey,

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

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

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

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

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

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

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

for the institutional posterior distribution $p_{PA}(\theta|X, Z)$, 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 $s(X, Y, \tilde{Z}, Z)$ an unbiased estimate of $s(X, Y, \tilde{Z})$.*

Corollary 4.2. *Let β be a parameter in the model $p_{SA}(Y|\theta, \tilde{Z})$ 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 $p(\beta|X, Y, \tilde{Z})$.*

Corollary 4.1 and Corollary 4.2 assert that Y, which is already a dependent variable in the secondary analyst's model $p_{SA}(Y|\theta, \tilde{Z})$, should also be an independent variable in the primary analyst's conditioning model $p_{PA}(\theta, Z)$, *in order that the standard PV methodology produces unbiased estimates* of $s(X, Y, \tilde{Z})$. While mathematically correct, this imposes a serious restriction on what the secondary analyst's model can be.

For ease of exposition, we consider the case in which $Z = (Y, \tilde{Z})$; as we see in Section 6 below, we can expect a similar behavior in the more general case $Z \supseteq (Y, \tilde{Z})$.

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

Note that $p(Y|\tilde{Z})$ is entirely determined by the observable relationship between Y and $\tilde{Z}$; consequently $p(Y|\theta, \tilde{Z})$ is completely 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

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

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

Theorem 4.1 imposes very strong constraints on the choices that the secondary analyst can make. If the secondary analyst's model $p_{SA}(Y|\theta, \tilde{Z})$ is the same as the one determined by Theorem 4.1 from $p(\theta|Y, \tilde{Z})$, then the PV methodology will ensure that $s(X, Y, \tilde{Z}, Z)$ is an unbiased estimate of $s(X, Y, \tilde{Z})$. Otherwise, $s(X, Y, \tilde{Z}, Z)$ is vulnerable to wrong-model bias, as an estimate of $s(X, Y, \tilde{Z})$.

In fact, the same problem exists, even if the survey institution's conditioning model contains a proxy U for Y, however poor, as the next proposition shows.

Corollary 4.3.

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

(b) If $Y \not\perp\!\!\!\perp U|\theta, \tilde{Z}$, *then* $p(\theta|U, \tilde{Z})$ *and* $p(U|\tilde{Z})$ *determine* $p(Y|\theta, \tilde{Z})$.

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,

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

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 $p(Y|\theta, \tilde{Z})$. □

Taken together, these results are pessimistic about the use of standard PV methodology to explore predictive inference using θ and other covariates: in order to ensure unbiased estimation of $s(X, Y, \tilde{Z})$ by $s(X, Y, \tilde{Z}, Z)$, the variable to be predicted, Y, must be in the institutional conditioning model, *and* the secondary analyst's model $p_{SA}(Y|\theta, \tilde{Z})$ must be the one implied by the survey institution's conditioning model. The bias when these conditions are violated can be substantial, as we will show below in Section 6, and may lead to incorrect scientific or policy conclusions.

As 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 can specify $p_{SA}(Y|\theta, \tilde{Z})$ correctly. 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 Section 4, we argued that the usual plausible values methodolgy, 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 Section 2, if $X \perp\!\!\!\perp \tilde{Z}, \beta|\theta$ and $\theta \perp\!\!\!\perp \beta|\tilde{Z}$,*