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) in particular 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.

Comparative studies of students' learning under different curricula, belonging to different social groups of interest, or residing in different countries, use estimated proficiency as the *dependent variable* in their analyses. Studies of 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 both of these cases the 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, perhaps surprisingly, 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) and in subsequent work of Mislevy 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, and Zelenak, 1999), and the structure of many other large-scale national and international surveys of education is similar (e.g., the National Adult Literacy Survey, NALS, Kirsch, et al, 2000; and Trends in International Mathematics and Science Study, TIMSS, Olson, Martin, and 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 actually 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) 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 Li, Oranje and Jiang, 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 PVs see Von Davier, Gonzalez, and Mislevy, 2009)

This raises 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, and perhaps counter-intuitively, any contrast (say, 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 Section 2 we review this

argument and see that releasing PVs based on an extremely large conditioning model is unbiased but perhaps statistically inefficient for the secondary analyst, when proficiency is a dependent variable.

When proficiency is an independent variable, however, we provide in Section 3 a disquieting result: the conditioning model must be carefully tailored to the secondary analysis that will be performed. When proficiency is used as an independent variable in a model for some other outcome (like wages), all other independent variables in the model *must* be in the conditioning model. However, the outcome variable or any variable associated with the outcome variable even after conditioning on proficiency *cannot* be used in the conditioning model. In short, we establish a kind of *Goldilocks Result*: when proficiency enters a secondary analysis as an independent variable, inclusion of either *too many* or *too few* covariates in the conditioning model can lead to biased effect estimates. The bias involves not only, say, the regression coefficient for proficiency in the model, but also regression coefficients for other predictors, whether they are latent or not.

We show two examples in Section 4 of the size and direction of the bias when violating our Goldilocks Result. In both of these examples, the structural model of interest is a linear regression model in which proficiency serves as an independent variable. In the first example, we calculate the bias of the regression coefficients when we include the outcome variable in the conditioning set. The first example assumes the measurement model has standard homoskedastic error as in a classical test theory (CTT) model. In the second example, we use simulations to show the bias under different conditioning models when we assume the measurement model is a standard IRT model with heteroskedastic error. It should be noted, however, that our results do not depend on the linear structure inherent in standard regression and SEM analyses, and as such are substantially more general.

An immediate consequence of our result 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 **??** we discuss workable machinery to do this.

2 When θ is the Response Variable

Let $X = [X_{ij}]$ be a (possibly incomplete) matrix of scored responses X_{ij} of survey subject i, $i = 1, \ldots N$, to cognitive item j, $j = 1, \ldots J$. Most commonly X_{ij} will be binary ($X_{ij} = 1$ for a correct response, and $X_{ij} = 0$ for an incorrect response), though sometimes X_{ij} will be an ordered or partially ordered categorical variable, a continuous variable, etc. Noninformative missingness is handled as usual by simply omitting terms from the appropriate likelihood; modeling informative missingness (e.g., Glas and Pimentel, 2008 and Holman and Glas, 2005) is beyond the scope of this paper. For ease of exposition we will treat X as complete, except where noted.

Let $\theta = (\theta_1, \ldots \theta_N)$ be the vector of latent proficiencies for the N survey respondents, and let

$$p(X|\theta, \gamma) = \prod_{i=1}^{N} p(X_{i1}, \ldots, X_{iJ}|\theta_i, \gamma) \tag{1}$$

be a probability model for X, given θ and possibly some other set of unknown parameters γ. Equation (1) is the usual "measurement model" that relates each subject i's responses $X_{i1}, \ldots, X_{iJ}$ to J items on a survey or cognitive assessment to the latent proficiency variable θ_i.

We assume the latent variables θ_i are sampled independently from some population distribution which may itself depend on covariates $Z = (Z_1, \ldots, Z_N)$ and other parameters τ,

$\theta_i \overset{indep}{\sim} \pi(\theta_i|Z_i,\tau)$, that is,

$$\theta \sim \pi(\theta|Z,\tau) \equiv \prod_{i=1}^{N} \pi(\theta_i|Z_i,\tau) \ . \tag{2}$$

The population distribution is formally equivalent to a prior distribution for θ_i and so a posterior distribution can be constructed for θ in the usual way,

$$p(\theta|X,Z,\tau,\gamma) \propto p(X|\theta,\gamma)\pi(\theta|Z,\tau) \ . \tag{3}$$

The covariates Z_i may include survey design variables, demographic or background variables, etc. Institutional PVs are simply random draws from the multivariate posterior $p(\theta|X,Z,\tau,\gamma)$, and can be used to compute means and other functionals of the posterior distribution, using Monte Carlo methods.

Secondary analyses involving θ as a dependent variable typically employ a conditional model of the form

$$p(\theta|\tilde{Z},\beta) \tag{4}$$

where $\tilde{Z}$ is a set of covariates and/or design variables relevant to the secondary analysis, and β is a set of parameters. As a simple case, consider the regression model

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

where $\tilde{Z}_i$ is a dummy variable for membership in some social group of interest. The goal is to estimate the effect β_1 of group membership, or more broadly to estimate the parameters β in the conditional model (4).

The central question is, what is the relationship between the set of covariates $\tilde{Z}$ in (4) and the covariates Z in (2)?

Mislevy's (1991) argument can be summarized as follows. The inference in the secondary analysis will be based on an estimator $\hat{\beta} = s(\theta, \tilde{Z})$ or some other function $s(\theta, \tilde{Z})$ designed for inference from the model (4). (In the simple linear regression illustration above, we might be interested in the least-squares estimator

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

for example.) Whatever the inference function $s(\theta, \tilde{Z})$, we cannot estimate it directly because θ is latent—unknown and unknowable (except in the limit of infinitely much data—see Junker, 1991; Ellis & Junker, 1997). Instead the best we can hope for is its expected value given the observable item responses X and covariates $\tilde{Z}$,

$$\begin{aligned} s(X, \tilde{Z}) &= E[s(\theta, \tilde{Z})|X, \tilde{Z}] \\ &= \int s(\theta, \tilde{Z}) p(\theta|X, \tilde{Z}, \tilde{\tau}, \gamma) d\theta \ , \end{aligned} \tag{5}$$

where $p(\theta|X, \tilde{Z}, \tilde{\tau}, \gamma)$ is the posterior density of θ as in (3), but using $\tilde{Z}$, with parameters $\tilde{\tau}$, rather than Z and τ.

Instead of $p(\theta|X, \tilde{Z}, \tilde{\tau}, \gamma)$, however, secondary analysts typically have only information about $p(\theta|X, Z, \tau, \gamma)$ from (3), usually in the form of institutionally-released PVs. Thus the calculation that is possible in secondary analysis is not (5), but rather

$$\begin{aligned} s(X, \tilde{Z}, Z) &= \int s(\theta, \tilde{Z}) p(\theta|X, Z, \tau, \gamma) d\theta \\ &= E[s(\theta, \tilde{Z})|X, Z] \ . \end{aligned} \tag{6}$$

Let us write, in a slight abuse of notation[1], $\tilde{Z} \subseteq Z$, if $\tilde{Z}$ is a subset or measurable function

[1]Indeed, if $\sigma(\tilde{Z})$ is the Borel sigma field generated by $\tilde{Z}$, and $\sigma(Z)$ is the Borel sigma field generated by Z, then by $\tilde{Z} \subseteq Z$ we mean precisely that $\sigma(\tilde{Z}) \subseteq \sigma(Z)$ as sigma fields.

of Z. If $\tilde{Z} \subseteq Z$, then

$$s(X, \tilde{Z}) = E[s(\theta, \tilde{Z})|X, \tilde{Z}] = E[E[s(\theta, \tilde{Z})|X, Z]|X, \tilde{Z}] = E[s(X, \tilde{Z}, Z)|X, \tilde{Z}] \ , \qquad (7)$$

with the second equality following from the "telescoping" property of conditional expectations (Billingsley, 1986 page 470); otherwise equality does not hold. Thus, $s(X, \tilde{Z}, Z)$ is an unbiased estimator of $s(X, \tilde{Z})$ if and only if $\tilde{Z} \subseteq Z$.

The degree of bias when equality in (7) does not hold, $s(X, \tilde{Z}) - E[s(X, \tilde{Z}, Z)|X, \tilde{Z}]$, depends on the mismatch between Z and $\tilde{Z}$, and we provide some examples in Section 4.

Even when equality holds in (7), the estimator $s(X, \tilde{Z}, Z)$ is, however, clearly inefficient: since the institution releasing PVs cannot know which $\tilde{Z}$ will be of interest to each secondary analyst, a very high-dimensional Z is used in (3), in hopes that any $\tilde{Z}$ of interest will be included directly or by proxy in Z. Given the same data, $p(\theta|X, \tilde{Z}, \tilde{\tau}, \gamma)$ could typically be estimated with much more precision than $p(\theta|X, Z, \tau, \gamma)$; for example whenever $\tilde{Z}$ is lower-dimensional than Z, $(\tilde{\tau}, \gamma)$ is likely to be lower-dimensional than (τ, γ), leading to more precise parameter estimates. In addition there is some computational inefficiency in calculating PVs for high-dimensional Z if only the lower dimensional $\tilde{Z}$ is needed; however this computational cost is born mostly by the institution releasing PVs.

What this approach may lack in statistical efficiency, it more than compensates in elegance and arithmetic consistency. Institutions that release PVs usually also base their primary reporting on the same PVs; thus all primary and secondary analyses are simply marginalizations, following (7), of the same set of PVs. Two analysts studying the same inference will draw numerically identical conclusions, if they take a valid approach to marginalizing over PVs, and all analyses will be logically and arithmetically consistent with the institution's primary reporting. Arguments over policy implications of secondary analyses need not devolve into disagreements about fundamental modeling assumptions, since

everyone is using the same model; numerical inconsistencies can only be due to arithmetic or rounding errors.

3 When θ is an Explanatory Variable

We continue with the same setup as in equations (1) and (2), but now we consider a case where a secondary analyst wants to estimate statistics in a model where θ is an independent variable,

$$p(Y|\theta, \tilde{Z}, \beta) \ . \tag{8}$$

For example, a secondary analyst may be interested in the regression model

$$Y_i = \beta_0 + \beta_1 \theta_i + \beta_2 \tilde{Z}_i + \varepsilon \ , \tag{9}$$

where Y is a response variable, θ is a latent proficiency that is now an explanatory variable, and $\tilde{Z}$ is another covariate explaining Y.

There is a modest and growing literature in the social sciences in which the test score serves as an independent variable (see Neal and Johnson, 1996; Venezky and Kaplan, 1998; and Ritter and Taylor, 2011 among many). Often, these analyses are regression-based and the cognitive test score is placed on the right-hand side of the regression equation as an explanatory variable, as in (9). In these analyses, there are two potential relationships of interest: First, analysts study the relationship between cognitive ability and some other outcome of interest, e.g., in research that evaluates the relationship between cognitive ability and likelihood of voting (see Venezkey and Kaplan, 1998). Second, researchers treat cognitive ability as a "control variable" in an analysis that focuses on the relationship between two (or more) other variables. For example, Neal and Johnson (1996) control for "pre-market human capital" to study black-white wage gaps and Ritter and Taylor (2011) do the same

for unemployment. In each of these examples, and in many others, researchers often use a test score (which is measured with error) as fixed data. Even in cases where researchers measure the error (as in Bollinger, 2003) the appropriate conditioning model is of interest.

Once again, since $s(\theta, Y, \tilde{Z})$ cannot be computed directly (θ is again unobservable); the best we can do is to calculate

$$\begin{aligned} s(X, Y, \tilde{Z}) &= E[s(\theta, Y, \tilde{Z})|X, \tilde{Z}] \\ &= \int s(\theta, Y, \tilde{Z}) p(\theta|X, \tilde{Z}, \tilde{\tau}, \gamma) d\theta \ , \end{aligned} \tag{10}$$

where again $p(\theta|X, \tilde{Z}, \tilde{\tau}, \gamma)$ is the posterior density for θ given X and $\tilde{Z}$, computed as in (3), with parameters $\tilde{\tau}$ and γ.

Again we consider what variables, Z should be included in a posterior density $p(\theta|X, Z, \tau, \gamma)$ from which to draw PVs for an institution to release, to help secondary analysts compute or estimate (10). That is, when will

$$\begin{aligned} s(X, Y, \tilde{Z}, Z) &= \int s(\theta, Y, \tilde{Z}) p(\theta|X, Z, \tau, \gamma) d\theta \\ &= E[s(\theta, Y, \tilde{Z})|X, Z] \end{aligned} \tag{11}$$

be an unbiased estimator for $s(X, Y, \tilde{Z})$?

The estimator $s(X, Y, \tilde{Z}, Z)$ will be unbiased for $s(X, Y, \tilde{Z})$ if

$$\begin{aligned} s(X, Y, \tilde{Z}) &= E[s(\theta, Y, \tilde{Z})|X, \tilde{Z}] \\ &\stackrel{?}{=} E[E[s(\theta, Y, \tilde{Z})|X, Z]|X, \tilde{Z}] \\ &= E[s(X, Y, \tilde{Z}, Z)|X, \tilde{Z}] \ , \end{aligned} \tag{12}$$

holds, and this depends on the telescoping property marked by "$\stackrel{?}{=}$" in (12) (as described in

Billingsley, 1986 page 470). As with (7), the equality holds if and only if $\tilde{Z} \subseteq Z$.

Thus it would seem that an institution releasing plausible values should again make Z in the conditioning model as large as possible, just as in Section 2. However, the new dependent variable Y, as well as θ in the role of independent variable, in (8) bears closer examination.

3.1 Including Y in the Conditioning Model

Let us consider what happens when the prior for θ, equation (2), is augmented to include Y. Instead of equation (2), we consider a new prior for θ as chosen by the primary researchers

$$\theta \sim \pi(\theta|Y, \tilde{Z}, \tau^*) \ . \tag{2*}$$

Intuitively, we see that when the prior for θ contains Y we have a circular argument: We want to determine the relationship between Y and θ using a θ that was calculated given that we knew Y.

More formally, given that Y and $\tilde{Z}$ are observed (and with enough data), we can determine $p(Y, \tilde{Z})$, the joint distribution of Y and $\tilde{Z}$. Using $p(Y, \tilde{Z})$, we integrate Y out of equation (2*) to gives us $\pi(\theta|\tilde{Z}, \tau^*)$. Elementary conditional density calculations then *force* a shape of the conditional distribution of Y given θ and $\tilde{Z}$, when the primary institution choses (2*) as the prior on θ:

$$\pi(Y|\theta, \tilde{Z}, \tau^*) \propto \frac{\pi(\theta|Y, \tilde{Z}, \tau^*)}{\pi(\theta|\tilde{Z}, \tau^*)}. \tag{13}$$

But, in order for the specification made by the secondary researcher in (8) to be logically consistent with the specification made by the primary institution (2*), it must be true that

$p(Y|\theta, \tilde{Z}, \beta) = \pi(Y|\theta, \tilde{Z}, \tau^*)$, which implies

$$p(Y|\theta, \tilde{Z}, \beta) \propto \frac{\pi(\theta|Y, \tilde{Z}, \tau^*)}{\pi(\theta|\tilde{Z}, \tau^*)} \ , \tag{14}$$

as a function of θ, for each possible set of Y's and $\tilde{Z}$'s. This reveals a new source of possible bias in developing the "institutional" Z: if $Y \subseteq Z$, then estimates of β may be biased due to a "wrong model bias." The extent of the bias depends on the extent that the secondary analyst's specification of equation (8) fails to obey the constraint in equation (14)—as it almost always would.

One might argue that we could pick equation (2^*) such that the prior chosen by the primary institute would match the model chosen by the secondary researcher. In order to specify $\pi(\theta|\tilde{Z}, \tau^*)$ such that it matches with the secondary analyst's model, we must know β. However, β or the relationship between θ and Y is precisely what we are trying to determine.

3.2 Including Additional Variables Other than Y in the Conditioning Model

We define a new variable U that is a subset of $Z \cap \tilde{Z}^C$. We now consider a prior for θ as chosen by the primary researchers that includes some U,

$$\theta \sim \pi(\theta|U, \tilde{Z}, \tau^{**}) \ . \tag{2**}$$

If a secondary analyst uses estimates of θ formed from the institutional prior in (2^{**}), in (8), their model for Y will now include U. This leads to an interesting question: For what cases does including U in the prior introduce a bias?

There are two cases. In the first case, we assume Y and U are conditionally independent of one another given θ and $\tilde{Z}$: $p(Y|\theta, \tilde{Z}, U, \beta) = p(Y|\theta, \tilde{Z}, \beta)$. In this case, the specification

of the prior chosen by the primary institution does not constrain the secondary analyst's chosen model. The model becomes

$$\begin{aligned}
\pi(\theta|U,\tilde{Z},\tau^{**})p(Y|\theta,\tilde{Z},\beta) &= \pi(\theta|U,\tilde{Z},\tau^{**})p(Y|\theta,\tilde{Z},\beta)\frac{p(Y|U,\theta,\tilde{Z},\beta)}{p(Y|U,\theta,\tilde{Z},\beta)} && (15)\\
&= \pi(\theta|U,\tilde{Z},\tau^{**})p(Y|U,\theta,\tilde{Z},\beta) && (16)\\
&= p(Y,\theta|U,\tilde{Z},\beta,\tau^{**}) && (17)\\
&= p(Y,\theta|\tilde{Z},\beta,\tau^{**}), && (18)
\end{aligned}$$

from which we can now determine β or any s that depends on Y, θ, and $\tilde{Z}$.

In the second case, Y are U are not conditionally independent of one another given θ and $\tilde{Z}$: $p(Y|\theta,\tilde{Z},U,\beta) \neq p(Y|\theta,\tilde{Z},\beta)$. In this case, we once again have "wrong model bias."

$$\begin{aligned}
\pi(\theta|U,\tilde{Z},\tau^{**})p(Y|\theta,\tilde{Z},\beta) &= \pi(\theta|U,\tilde{Z},\tau^{**})p(Y|\theta,\tilde{Z},\beta)\frac{p(Y|U,\theta,\tilde{Z},\beta)}{p(Y|U,\theta,\tilde{Z},\beta)} && (19)\\
&= p(Y,\theta|U,\tilde{Z},\beta)\frac{p(Y|\theta,\tilde{Z})}{p(Y|U,\theta,\tilde{Z})} && (20)\\
&\neq p(Y,\theta|\tilde{Z},\beta) && (21)
\end{aligned}$$

In the case where, U and Y are conditionally independent of one another, U is not "informative" for Y outside of U's relationship with θ and $\tilde{Z}$. We show in Section 4 that when U is not informative for Y, estimates of $s(Y,X,\theta)$ are *more* accurate, because U acts as a kind of "instrument" or "exclusion restriction" on θ providing better estimates of θ and thus reducing measurement error.

Thus we arrive at a kind of "Goldilocks condition" for an institutional Z:

- Z must contain $\tilde{Z}$ in the secondary analyst's specification (8);
- Z must *not* contain Y in equation (8); and
- Z must *not* contain any variable that is informative for Y except through $\tilde{Z}$ and θ in

equation (8).

In the next section we consider several analytic and simulation examples that illustrate the extent and direction of the bias when $\tilde{Z} \not\subseteq Z$, when $Y \subseteq Z$, and/or when some $U \subseteq Z$. Our examples suggest there is no "universal set" of plausible values that an institution can release if θ will be used as an independent variable by secondary researchers. Instead, secondary analysts must specify a prior distribution (2) customized to their particular prediction model (8), such that the prior (2) conditions on all of $\tilde{Z}$ in (8), but does not condition on Y or any variable that is informative for Y outside of its relationship with θ and $\tilde{Z}$ in (8).

4 Examples

4.1 Example 1: A Normal-Normal-Normal Model

We show here the bias that results for a specific case when θ is an independent variable and Y is in the conditioning model. We specify a simple structural equations model such that the functional form of the structural model, the measurement model, and the conditioning model are all normally distributed. The choice of normality for all levels in the hierarchical model forces a closed analytic form of the posterior distribution of the parameters of interest. Our model is

$$
\begin{aligned}
Y_i^*, |\theta_i, Z_i^*, \beta, \sigma^2 &\overset{\text{indep}}{\sim} N(\beta_0 + \beta_1\theta_i + \beta_2 Z_i^*, \sigma^2) \\
\theta_i | Z_i^*, Y_i^*, \tau^2 &\overset{\text{i.i.d.}}{\sim} N(\alpha_1 Z_i + \alpha_2 Y_i^*, \tau^2_{Z^*,Y^*}) \\
X_i | \theta_i, \lambda^2 &\overset{\text{i.i.d.}}{\sim} N(\theta_i, \lambda^2) \\
\beta_2 &\sim N(\gamma, \xi_2^2).
\end{aligned}
$$

We choose a particularly simple case of the structural model, namely when $\tilde{Z}_i$ includes