Measurement Error Models for Latent Trait Estimates: Interpretation and Inference

Adam VanIwaarden

University of Colorado

Abstract

An analysis of the extent and impact of measurement error, and correction for it, requires a defensible error model. Contrary to the convention of assuming measurement errors conform to the classical model, which is pervasive in the social sciences, it is suggested here that the Berkson error model is appropriate when parameter estimates are used as measurements of an unobservable entity (i.e. latent trait estimates). An introduction to the Berkson and Classical error models is provided along with a description of the potential problems the inclusion of latent trait estimates pose to statistical inference. The potential use and limitation of correction methods is also considered. Estimates of academic ability commonly produced through educational assessment are used as a pertinent example throughout.

Measurements of observable, physical entities are always subject to error and this is equally true when parameter estimates are used as "measurements" of unobservable, latent traits. However, *important differences between parameter estimates and physical observations (measures) suggest that they require separate modes to accurately describe measurement error, or uncertainty*. This assertion does not appear to be widely recognized in the social science literature. Although the distinction between error types is increasingly common in fields such as health sciences (e.g. many recent studies in radiation dosimetry, as showcased in Ron and Hoffman, 1999 and Schafer and Gilbert, 2006), there are few examples of non-classical error types used or discussed in the social sciences.

Highlighting the distinctions between error models is important because assertions about the impact of one error type do not apply automatically to another. For example, the bias in parameter estimates caused when error-prone variables are used as covariates in regression models is only true for classical errors (Carroll et al., 2006). Error correction methods that are consistent for one type of error are not for others (Carroll et al., 2006; Delaigle, 2007), and some error types are harder to account for than others (Heid et al., 2004).

In this paper I argue that latent trait parameter estimates are most appropriately described as a "Berkson" type error. In the first section I detail the commonalities and differences between the Berkson and classical definitions of error and their associated mathematical models. I then present measures of academic "ability" as an example of a latent trait commonly used in education and other social sciences with particular detail given to their errors of measurement. Finally, I discuss the how error model assumptions can impact inferences made when error prone variables are included in statistical models.

Error defined and two prominent error models

Error is defined as the deviation of a measured value from its "true" value. An *error model* then refers the mathematical formulation of this uncertainty, highlighting the relationship between the fallible measurement and truth. Error models are an interpretation of uncertainty, and the assumptions made in them are critical to the ability to 1) correct for the influence of error and 2) make viable inferences based on error-prone measures. One of the key assumptions pertains to the structure. For example, it may be assumed that error is random (i.e. zero expectation) and additive, with identical variance and distribution for all values (*homo*scedastic). Error models implicitly contain assumptions about the statistical independence of the error. This is the primary distinction between error types. The *classical* type error is independent of the true variable, whereas *Berkson* type error is independent of the observed (or estimated) variable.

In general, classical errors arise when direct, individual observations are made using an imperfect measurement instrument. Replications of this measurement process would result in a distribution of observed values centered around the true value. Berkson errors, on the other hand, result when an observation or parameter estimate from a subpopulation is assigned to an individual in place of a direct observation of the value (Schafer & Gilbert, 2006). Group identity or some other common factor may determine the assignment of these values, or they may be model-based estimates given common values of some observed proxy variable.

For the mathematical formulation of these two error models, let t be the theoretical "true" value of some unobservable variable of interest, x represents a measurement of it and e is a random variable representing the difference between the true and measured values. The classical measurement error (CME) model is defined as

$x_{j,r} = t_j + e_r$ (or "observed = true + measurement error").

The Berkson measurement error (BME) model, on the other hand, is

$t_{j,k} = x_j - e_k$ (or "true = estimated - individual peculiarity").

The errors, e, have different interpretations under the two models (Schafer & Gilbert, 2006). The CME model supposes a situation in which there are (ideally) a series of r repeated observations for individual j whose unobservable true value is equal to t_i. We would observe r unique values of x, which are distributed around their expected score (with deviations based on the r values of e_r). That is, in the basic CME model the random process is located at the level of the individual and errors are interpreted as random measurement inaccuracies. Estimates of measurement error variance can be used be used to predict the distribution of obtained scores based on the unobserved true score, but it is not the error variance of the estimate for predicting true score from observed (Harvill, 1991). If x is measured with unbiased error, $\bar{x}_r$ is its mean from r observations and

$$\mathbb{E}(\bar{x}_r \mid t = t_j) = t_j$$

where $\mathbb{E}(\bar{x}_r)$ is the expected value ($\mathbb{E}$ is the expectation operator) of $\bar{x}_r$ if the whole process could be repeated a large number of times (Berkson, 1950). In this model the error is independent of true score, or at least $\mathbb{E}(e_r \mid x_r) = 0$.

On the other hand, the Berkson model suggests a situation in which k individuals are assigned the same value of $x = x_i$, but their individual true value, t_{jk}, differs from that value by e_k. In this scenario we have a distribution of true values, t_j, which vary around the fixed group value, x_j (Berkson, 1950, Schafer & Gilbert, 2006). The random process is located at the level of the subgroup and error is interpreted as individual differences from the group-level estimate. The error variance can be used to predict a range of feasible true values given the observed estimate. If e is an unbiased error, $\bar{t}_k$ is the true average value of the subpopulation of k individuals and

$$\mathbb{E}(\bar{t}_k \mid x = x_j) = x_j$$

where $\mathbb{E}(\bar{t}_k)$ is the expected value of $\bar{t}_k$ approached when the process is repeated (Berkson, 1950). In this model the error is independent of the observed estimate, or $\mathbb{E}(e_k \mid t_k) = 0$.

Latent trait estimates of academic ability

Research questions and policy concerns in education often focus on the latent trait generally referred to as *academic ability*. For example, when we ask whether a student has "grown" after a year of schooling the question is generally regarding changes in the underlying ability trait and not on mere changes in the number of test items answered correctly. However, test scores remain an important entity here as they serve as a predictor variable in the estimation of unobservable ability. Modern test theories have proposed various mathematical models to describe the functional relationship between test scores (or individual item responses) and ability. Given that various model assumptions hold, the estimation of the latent trait is possible. These distinct, yet complementary, processes of 1) observing item responses and 2) producing trait estimates from them lead to classical and Berkson errors respectively.

Both the size and type of measurement errors are determined in part by the mathematical model used to summarize test item responses as "scores." The greater the complexity of the *operationally defined predictor* (a measurable and valid surrogate for the predictor of interest) the more difficult it is to measure and model its error (Heid et al., 2004). Since a sum score consists of directly observed responses, a simple classical error model can accurately describe its measurement error. An estimate of ability, on the other hand, uses item responses as predictors rather than as observations. Uncertainty in estimates such as this is more accurately described by the Berkson error model (Delaigle, 2007; Heid et al., 2004).

There are numerous specifications of mathematical models for the estimation of academic ability, but they can be generalized as follows. An item response vector **U** is

postulated as related to ability, θ, by a monotonically increasing function g, with parameters $\boldsymbol{\phi}$ such that

$$\bar{\theta}_j = g(\mathbf{U_j}|\boldsymbol{\phi}).$$

Here $\bar{\theta}_j$ is the *true* typical value of ability for a group of students with the same observed item response vector $\mathbf{U}_j$ on a test. Although the function g can represent any latent trait model used to estimate ability (e.g. factor analysis, structural equation models), I focus on the family of logistic item response models commonly used in Item Response Theory (IRT).

In the general Berkson error model presented above, we would replace an individuals' true ability, θ_j, with the true group-level typical value of ability, $\bar{\theta}_j$

$$\theta_j = \bar{\theta}_j + E_{Bj}.$$

In practice, the true value of $\bar{\theta}_j$ is replaced with an uncertain estimator of it, $\hat{\theta}_j$,

$$\hat{\theta}_j = g(\tilde{\mathbf{U}}_j|\hat{\boldsymbol{\phi}}).$$

I focus on maximum likelihood estimators (MLE) of the ability parameter, although the results generalize to similar estimators.

Individual level variation in item responses over replications of the measurement procedure would be another potential source of error. Variation of this sort would result in classical measurement errors because a student's observed item response vector $\tilde{\mathbf{U}}_j$ is itself random (not fixed as implied in the model) and varies from the true response vector, $\mathbf{U}_j$. This translates to the basic classical model presented above: $\tilde{\mathbf{U}}_j = \mathbf{U}_j + \mathbf{E}_{Cj}$. This type of error is not considered in either the IRT model or the Berkson error model. Although a distinct account of both is necessary in order to fully account for measurement error issues, I focus solely on the Berkson error introduced in the latent trait estimation process in this study. More detailed analyses that account for the potential mixed error model will be the subject of future research.

The non-classical error models described here are rarely considered in social science research. A search of the ERIC database returns no results for the term "Berkson error," suggesting that even this relatively simple error model has not been addressed in the educational context. Furthermore, recent research on the impacts of measurement error and applications of statistical corrections that specifically utilize latent trait estimates assume an additive, classical error structure (c.f. Shang, 2012 and AIR, 2012 for academic and policy applications respectively). In the following sections I provide a detailed description of the IRT mathematical models, with particular attention given to the estimation of error variances in order to endorse the use of the Berkson error model instead of the classical error model when latent trait estimates are the operationally defined predictor.

Ability estimation and uncertainty in Item Response Theory

The "Rasch" model is a relatively simple estimator of academic ability. This IRT model begins with the expectation of the response, u, to an item, i, is related to the latent ability, θ, through the functional form

$$h\left[\mathbb{E}\left(u_{ij}\right)\right] = b_i + \theta_j,$$

where b_i represents item difficulty and h is a link function. The expectation refers to the proportion of students with *true* ability θ_j expected to correctly answer item i (Borsboom, 2005). The logit transformation is commonly used for the link function h. Here the natural logarithm for the odds of a correct response leads to a model with the form

$$\ln\left[\frac{\mathbb{E}\left(u_{ij}\right)}{1-\mathbb{E}\left(u_{ij}\right)}\right] = b_i + \theta_j.$$

This can also be expressed as the probability of a correct response given ability, known generically as the *probability density function* (PDF) or as the *item response function* (IRF) in IRT terminology.

$$\Pr(u_{ij} = 1|\theta_j, b_i) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)}.$$

The PDF for the entire item response vector for a student with given ability θ and known item parameter(s) $\boldsymbol{\chi}$ can then be expressed as the multiplication of item PDFs,

$$\Pr(\mathbf{u}_j = (u_1, u_2, \ldots u_n)|\theta; \boldsymbol{\chi}) = \Pr_1(u_1 = 1|\theta)\ \Pr_2(u_2 = 1|\theta) \ldots \Pr_n(u_n = 1|\theta).$$

This suggests that certain item response vectors are more probable than others. However, in the present situation, we have already observed the item responses. The task at hand is to estimate the unknown ability parameter based on the assumed model, estimated item parameters and the observed response data. There are multiple different values of θ that could potentially produce an observed response vector, but there is only one value that makes it the *most likely* and this value is the "maximum likelihood estimate." We reverse the roles of the response vector $\mathbf{u}$ and θ to define the likelihood function, L,

$$L(\boldsymbol{\theta}|\mathbf{u}; \boldsymbol{\chi}) = \Pr(\mathbf{u}|\boldsymbol{\theta}; \boldsymbol{\chi})$$

to solve the estimation problem (Greene, 2007; Myung, 2003). The difference between these two functions is more than just a semantic one. The PDF is a function of the observed data with θ fixed, whereas L is a function of the ability estimate given fixed data (Myung, 2003). Defining the likelihood function on the ability scale in this way corresponds with the Berkson error model, where observed proxy values are considered fixed and the distribution of interest is that of possible true values that vary around the estimate based on the fixed proxy observations.

The estimator $\hat{\theta}_{MLE}$ has the desirable maximum likelihood properties. It is consistent, so it converges in probability the true value of θ and has an asymptotically normal distribution. These asymptotic results are with respect to the increasing number of items on a test, not the number of students tested (Lord, 1980; Kolen, 2011). The conditional variance of the estimator is given as the inverse of the expected Fisher information (or "test information", denoted as $TI\{\theta\}$). This variance is also the formal measure of uncertainty in IRT parameter estimates, referred to as the conditional standard error of measurement (CSEM). So we have

$$\hat{\theta} \sim N(\theta, TI\{\theta\}^{-1})$$
$$E \sim N(0, TI\{\theta\}^{-1})$$

The test information is computed as

$$TI\{\theta\} = \mathrm{Var}(\hat{\theta}) = \left[\sum_{i=1}^{n} \frac{Pr_i'^2}{Pr_i(1 - Pr_i)}\right]^{-1}$$

(Lord, 1980), where Pr_i' is the first derivative of the PDF. This can be simplified as

$$\mathrm{Var}(\hat{\theta}) = \left[\sum_{i=1}^{n} P_i(\theta)(1 - P_i(\theta))\right]^{-1}$$

in the Rasch model (Embretson & Reise, 2000; Doran, 2005). This is simply the inverse of the sum of the binomial variances for the probability of a correct response. Note that the variance of the estimator is at its lowest when the binomial variance is greatest.

Lord (1980) shows that this compound binomial in the denominator of the final equation is an estimate of measurement error for the raw sum scores, and so the information and error variance for ability estimates are the inverse of the observed sum score information and error variance. In other words, the observable number correct

> "true score ξ can be estimated very accurately for such [high scoring] examinees: it is close to n. Their ability θ cannot be estimated accurately; however: We know that their θ is high without knowing how high. This situation is mirrored by the

> fact that $I\{\xi, x\}$ is very large for such examinees, whereas $I\{\theta, x\}$ is near zero. The reader should understand these conclusions if he is to make proper use of information functions (or of standard errors of measurement)" (Lord, 1980, p 90).

A close look at the last equation shows that this estimate of uncertainty is purely a function of the unobservable ability parameter. The observed data does not factor into the calculation. It is strictly describing the precision of the ability parameter estimate. This type of error variance is fundamentally different than those pertaining to test score consistency (Doran, 2005). Yet there are examples of both academic research and policy implementation that specifically address measurement error when using latent trait estimates in which these two error types are treated interchangeably (cf. Shang, 2012; Air, 2012 respectively). As has been described here, the estimator variance corresponds to a Berkson error model, with error independent of the observed estimate, which describes the distribution of the possible true parameter values that could have generated the fixed, observed data (given the model assumptions hold).

Practical implications for error model assumptions

To this point the arguments made in favor of situating uncertainty in latent trait estimates within the Berkson paradigm have been primarily academic, focusing on the theoretical alignment between the measurement and error models. I now provide a more practical argument, using the results from a simple simulation study to show how the assumptions a researcher makes about the error model can play out in the statistical inferences one might make when dealing with error prone data. In this simulation study I deal with random, non-differential, and additive errors of both the Berkson and classical types. I vary the size of the measurement errors and review both homo- and heteroscedastic error structures.

For the classical error model, "true" prior scores (x) are first created by drawing 500 observations from a random normal distribution (mean = 0 and standard deviation = 1). For the homoscedastic, "small" measurement error simulations a similar random, normal value is drawn with a standard deviation of 0.5, and the "large" measurement error with a standard deviation of 1.0. These simulated error values are added to the "true" values to obtain the simulated observed values.

The same small and large error values used for the classical error model are also used for the Berkson error model simulations. However, the role of the observed values and true values are switched in this simulation. As per the Berkson model, we begin by simulating 500 observed values (again drawn from a random normal distribution with mean = 0 and standard deviation = 1) and then add the measurement errors to obtain the simulated true values.

This process is then repeated for the heteroscedastic errors. Here the errors are simulated such that the error variance is greater at the lower and upper ends of the score scale (creating a U shaped conditional distribution). In both the homoscedastic and heteroscedastic simulations, the dependent variable is simulated at the true value of x plus a regression error: $y = x + e_{regression}$. The regression error is a random normal variate with mean = 0 and standard deviation = 0.5.

Bias and Mean Squared Error (MSE) are statistics typically used to compare the performance of two competing estimators. Bias is defined the difference between the estimated and the true parameter averaged over simulation replications, and MSE is the squared difference between the estimated and the true parameter averaged over simulation replications. Here the parameter of interest is the "true" slope of the regression line, which is set equal to 1. The following statistics are derived to compare the simple linear regression results based on a single classical and Berkson error-prone covariate.

$$\text{MSE} = \frac{1}{R}\sum_{r=1}^{R}(\hat{\beta}_{1,r} - 1)^2$$

Where $r = 1, 2, \ldots, R$ indicates 10,000 simulation replications, $\hat{\beta}_{1,r}$ denotes the estimated slope parameter for replication r, and 1 denotes the "true" slope parameter.

$$\text{Bias} = \frac{1}{R}\sum_{r=1}^{R}(\hat{\beta}_{1,r} - 1)$$

Table 1 presents the results of the simulation study where the size of the measurement error is relatively small.

Table 1. Mean Squared Error and Bias over 10,000 Simulations: Small error

	Classical Error Model		Berkson Error Model	
	MSE	Bias	MSE	Bias
Homoscedastic Error	0.040	-0.200	0.001	0.000
Heteroscedastic Error	0.060	-0.244	0.001	0.000

Table 2 presents the results of the simulation study where the size of the measurement error is relatively large.

Table 2. Mean Squared Error and Bias over 10,000 Simulations: Large error

	Classical Error Model		Berkson Error Model	
	Total MSE	Mean Bias	Total MSE	Mean Bias
Homoscedastic Error	0.250	-0.500	0.003	0.000
Heteroscedastic Error	0.230	-0.480	0.003	0.001

The results of this short and simple simulation study suggest that the correct identification of an error model is critical when making inferences based on error prone data. As is typically described in most textbook accounts of the classical error model, we see increasingly biased estimates of the regression slope parameter when *classical* error prone a variable is used as a covariate. However, we also see that this is not the case when latent trait estimates containing an identical amount of uncertainty are used. Here the main concern is not the parameter bias (or lack thereof), but of the additional uncertainty in the parameter estimate that results from this type of measurement error. This additional uncertainty is potentially as

problematic as bias, yet is more difficult to correct for given the difficulty in accurately assessing the amount of Berkson error that may be present in a variable (Heid et al., 2004).

Ignoring classical measurement error when present would lead one to make spuriously inferences (biased downwards). This is not the case in the Berkson model, although the uncertainty in the estimates is likely to be underestimated. The potential to over-correct for errors is also suggested here. If one were to assume that the error followed the classical model, when in fact it was Berkson, and applied a consistent correction method for classical error, the "corrected" estimates would likely be biased, introducing problems that did not previously exist (Carroll et al., 2006).

Discussion

To be completed prior to conference Nov. 1…

Figure 1. Classical error model simulations of different size and structure

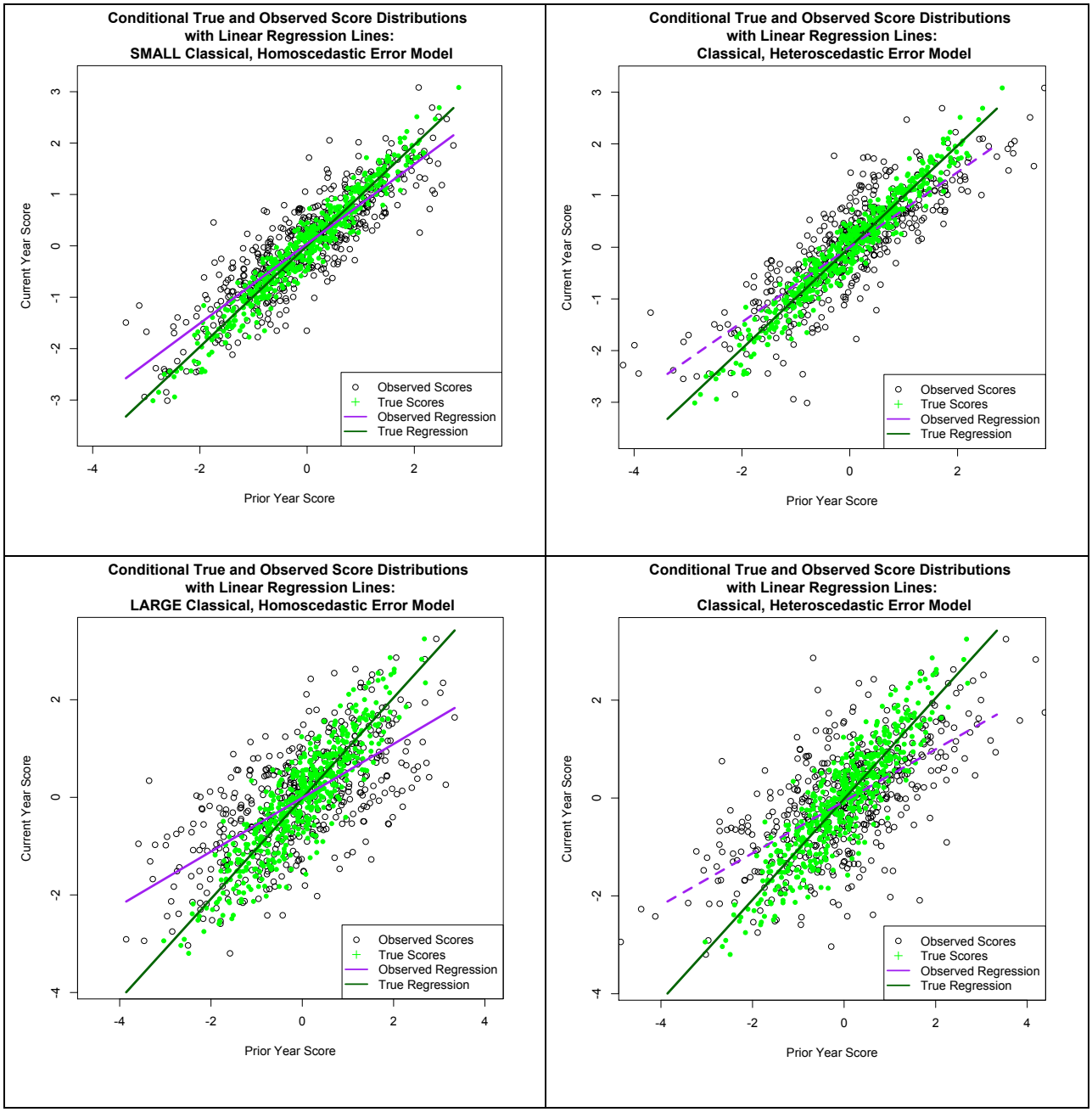
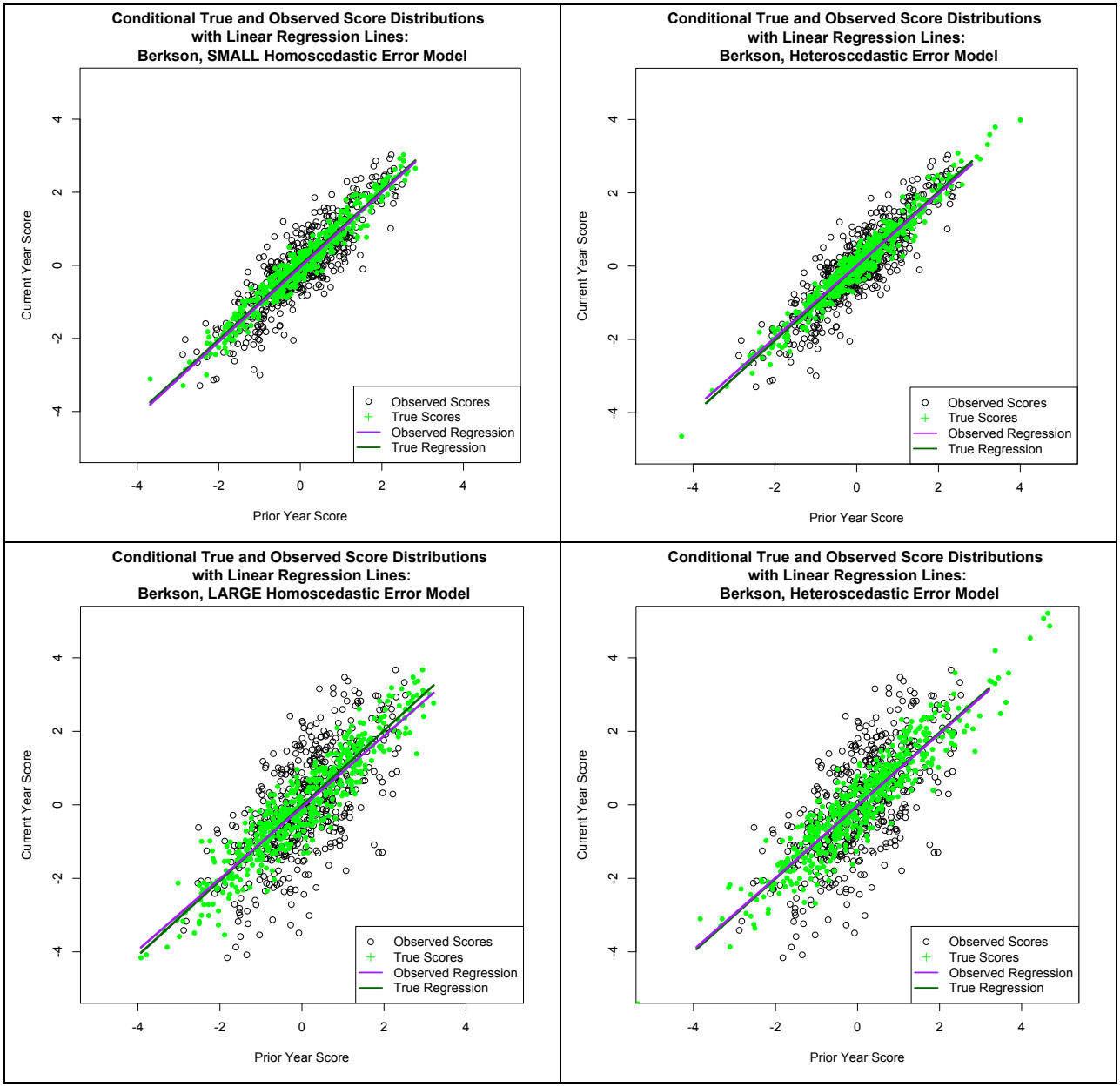


Figure 2. Berkson error model simulations of different size and structure



References

American Institutes for Research. (2012). 2011-2012 Growth Model for Educator Evaluation Technical Report: FINAL. Accessed December 20, 2012: http://engageny.org/sites/default/files/resource/attachments/growth-model-11-12-air-technical-report_0.pdf

Berkson, J. (1950). Are There Two Regressions? *Journal of the American Statistical Association, 45*(250), 164-180.

Borsboom, D. (2005). Measuring the Mind: Conceptual Issues in Contemporary Psychometrics. New York: Cambridge.

Buonaccorsi, J. and Lin, C. (2002). Berkson measurement error in designed repeated measures studies with random coefficients. *Journal of Statistical Planning and Inference*, 104, 53-72.

Buzas, J., Tosteson, T. and Stefanski L. (2003). Measurement Error. Institute of Statistics Mimeo Series No. 2544. Accessed March 1, 2012: http://www.biostat.washington.edu/~yanez/b578D/materials/mimeo_series_mem_version_master_file.pdf

Carroll, R., Ruppert, D., Stefanski, L., and Crainiceanu, C. (2006): Measurement Error in Nonlinear Models. Chapman & Hall, Boca Raton, FL.

Carroll, R., Delaigle, A. and Hall, P. (2007). Non-parametric regression estimation from data contaminated by a mixture of Berkson and Classical errors. *Journal of the Royal Statistical Society B*, *69*(5), 859-878.

Delaigle, A. (2007). Nonparametric density estimation from data with a mixture of Berkson and Classical errors. *Canadian Journal of Statistics*, *35*(1), 89-104.

Delaigle, A., Hall, P. and Qiu, P. (2006). Nonparametric methods for solving the Berkson errors-in-variables problem. *Journal of the Royal Statistical Society B*, *68*(2), 201–220.

Doran, H. (2005). The Information Function for the One-Parameter Logistic Model: Is it Reliability? *Educational and Psychological Measurement, 65*, 665-675.

Embretson, S., and Reise, S. (2000). Item response theory for psychologists. Hillsdale, NJ: Lawrence Erlbaum.

Fischer, G. and Molenaar, I. (1995). Rasch Models: Foundations, Recent Developments and Applications. New York: Springer-Verlag.

Greene, W. (2007). Econometric Analysis (6 Ed.). Upper Saddle River, NJ: Prentice Hall.

Heid, I., Kuchenhoff, H., Miles, J., Kreienbrock, L. and Wichmann, H. (2004). Two dimensions of measurement error: Classical and Berkson error in residential radon exposure assessment. *Journal of Exposure Analysis and Environmental Epidemiology 14*, 365–377.

Kolen, M., Zeng, L., and Hanson, B. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement*, *33*(2), 129–140.

Kolen, M. and Lee, W. (2011). Psychometric Properties of Raw and Scale Scores on Mixed-Format Tests. *Educational Measurement: Issues and Practice, 30*(2), 15-24.

Lee, W., Brennan, R. L., and Kolen, M. (2000). Estimators of conditional scale-score standard errors of measurement: A simulation study. *Journal of Educational Measurement, 37*(1), 1–20.

Lord, F. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.

Myung, J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology,* 47, 90-100.

Ron, E., and Hoffman, F.O., (eds.). 1999. Uncertainties in Radiation Dosimetry and Their Impact on Dose-Response Analysis. Proceedings of a Workshop at the National Cancer Institute, September 3-5, 1997. NIH Publication No. 99-4541. National Cancer Institute, Bethesda, MD.

Schafer, D. and Gilbert, S. (2006). Some Statistical Implications of Dose Uncertainty in Radiation Dose-Response Analyses. *Radiation Research, 166*(1, Part 2), 303-312.

Shang, Y. (2012). Measurement error adjustment using the SIMEX method: An application to student growth percentiles. *Journal of Educational Measurement*, 49(4), 1-20.