Goldstein, H., Haynes, M., Leckie, G., & Tran, P. (2020). Estimating reliability statistics and measurement error variances using instrumental variables with longitudinal data. *Longitudinal and Life Course Studies*, *11*(3), 289-306. https://doi.org/10.1332/175795920X15844303873216

# Estimating reliability statistics and measurement error variances using instrumental variables with longitudinal data.

by

Harvey Goldstein, University of Bristol (h.goldstein@bristol.ac.uk)
Michele Haynes, Australian Catholic University (michele.haynes@acu.edu.au)
George Leckie, University of Bristol (g.leckie@bristol.ac.uk)
Phuong Tran, Australian Catholic University, (phuongtran.12t@gmail.com)

## Abstract:

The presence of randomly distributed measurement errors in scale scores such as those used in educational and behavioural assessments implies that careful adjustments are required to statistical model estimation procedures if inferences are required for 'true' as opposed to 'observed' relationships. In many cases this requires the use of external values for 'reliability' statistics or 'measurement error variances' which may be provided by a test constructor or else inferred or estimated by the data analyst. Popular measures are those described as 'internal consistency' estimates and sometimes other measures based on data grouping. All such measures, however, make particular assumptions that may be questionable but are often not examined. In this paper we focus on scaled scores derived from aggregating a set of indicators, and set out a general methodological framework for exploring different ways of estimating reliability statistics and measurement error variances, critiquing certain approaches and suggesting more satisfactory methods in the presence of longitudinal data. In particular, we explore the assumption of local (conditional) item response independence and show how a failure of this assumption can lead to biased estimates in statistical models using scaled scores as explanatory variables. We illustrate our methods using a large longitudinal dataset of mathematics test scores from Queensland, Australia.

**Key words:**

Reliability, longitudinal data, instrumental variables

# 1. Introduction

In practice, one or more explanatory variables in linear and generalised linear models are often measured with error. It is well known that if inference is required about the relationship for the underlying 'true' values, then using the observed 'error-full' values will generally lead to biased and inconsistent inferences. A number of standard procedures for handling such data have been suggested (Fuller, 2006), as well as more advanced procedures, such as the SIMEX method (Cook and Stefanski, 1994, Carroll et al., 1996) and more recent Bayesian procedures (see for example Goldstein et al., 2017). For all of these procedures it is assumed that a good estimate of the measurement error distribution is available. In this paper we consider the case where the explanatory variables subject to measurement error are scaled scores derived from a set of indicators, for example a set of binary correct/incorrect responses in an educational test or a set of agree/disagree questions in an attitude scale. In many educational and behavioural models such error-full explanatory variables are used, based on rating scales or test scores, for example, in the case of value-added school performance models (Leckie and Goldstein, 2019). We focus on such cases where the explanatory variable is the sum of a set of, typically binary, indicators.

For continuous explanatory variables, or pseudo-continuous variables treated as continuous, a major issue in all these procedures is obtaining a satisfactory estimate of the reliability or measurement error variance and this paper is devoted to a discussion of different approaches with recommendations of what would be appropriate in the context of longitudinal data. We consider in detail the linear regression model where explanatory variables include variables with measurement error and where we also have variables measured at different occasions on the same individual units. Our examples are largely drawn from education but are readily applicable to other areas. In the next section we formally introduce the measurement error model, and for completeness, briefly discuss how knowledge of the measurement error distribution enables consistent estimation of the model parameters.

We note that our proposed procedure, despite our focus on scaled scores, is generally applicable, as described in later sections, to scores or ordered classifications however derived. A general IV approach to estimation with measurement errors is given by Meier et

al. (2017) but they do not study the specific case of scaled scores, and the exposition in the present paper is original

## 2. A basic measurement error model

We begin by assuming a simple normal linear regression model where the measurement error model can be written as

$$x_i = X_i + m_i \quad . \tag{1}$$

Here, capital letters refer to the true values, lower case letters refer to the observed values, and $m_i$ denotes the measurement error for the explanatory variable $X_i$. We also assume $y_i = Y_i$, that is, and without loss of generality, we assume no measurement error in the response. We also make the standard assumption used in measurement error models, that $X_i, m_i$ are independent of each other and the $m_i$ are mutually independent.

The model of interest (MOI) is

$$Y_i|X_i = \beta_0 + \beta_1 X_i + e_X \tag{2}$$

whereas what we observe is

$$Y_i|x_i = \alpha_0 + \alpha_1 x_i + e_x. \tag{3}$$

The subscript $i$ has been dropped in some cases for ease of expression.

Write $m_i \sim N(0, \sigma_m^2)$, $X_i \sim N(\mu, \sigma_X^2)$, $R = \sigma_X^2 \Big/ (\sigma_X^2 + \sigma_m^2)$

and in this simple model we have $R = \alpha_1/\beta_1$, where *R* is known as the 'reliability' of $x_i$. Clearly, if we have a good estimate for $R$ or $\sigma_m^2$ we can use (3), based on the observed data, to obtain a consistent estimator of $\beta_1$. As discussed above, there is an extensive literature about such 'reliability' or more accurately, measurement error, corrections and the model specified by (1)-(3) is known as the classical measurement error model. A major problem, however, remains in that it is not always straightforward to obtain good estimates for $R$ or $\sigma_m^2$. Our discussion focusses on obtaining estimates for these quantities. First, however, we briefly discuss the target population.

We note that the reliability depends upon both the measurement error variance and the population distribution of the true (and observed) values. Either or both properties, and therefore the reliability, may vary by subpopulation where a subpopulation is defined as the target of interest, for example females. Even when the reliability does not change across

subgroups, the measurement error distribution will do so if the observed variances differ. Thus, when fitting models to data with heterogeneous measurement errors this should be incorporated, else a failure to accommodate these can itself lead to biases. Goldstein et al (2017) discuss how this may be done and we also provide an elaboration below.

For generalised linear models where the response is, for example, binary, estimation will generally be more complex but for our purposes no new features are introduced. In particular for Bayesian generalised linear models, we can simply introduce an extra step when fitting the model of interest in the appropriate MCMC algorithm to sample, conditional on the value of the response, from an underlying latent normal distribution so that the modified response is normal (Goldstein et al., 2017).

## 3. Estimating the measurement error variance

Where we have independent replications of the measurements with errors, it is generally possible to obtain direct estimates of the measurement error variance by utilising the variation between replicates. For example, replicate measurements of baby length at a health clinic. In the case of scaled scores, however, especially with cognitive measurements, this will not be possible largely due to familiarity, memory or learning effects. We now examine two different approaches to this problem.

### 3.1 Internal consistency estimation methods

This approach is used extensively in psychology and education for scaled scores where a variable is scored by summing a set of constituent parts. Thus, an attainment score might consist of a set of binary correct/incorrect items with each scored 1 if correct and 0 if incorrect. The item scores are typically summed to form a total score. Reliabilities are then estimated using what might be termed a pseudo-replication method as follows (Lord and Novick, 2008).

Consider a variable derived as follows:

$$x_i = \sum_{j=1}^{k} p_{ij} \qquad (4)$$

where we assume for simplicity that $p_{ij}$ takes the values (0,1) for a *k*-item test. If we divide the test items at random into two (approximately) equal groups and we assume that, for any given testee, the response to one item is independent of the response to any other

item, an assumption of conditional or local independence, then we can treat the total scores from each group as an independent replicate and hence obtain an estimate for the 'half test', between-replicate covariance. Thus, for the whole test score, an equivalent estimate of the measurement covariance would simply be four times this value. If we now take all possible such splits, this leads to the standard Kuder-Richardson (KR20) (a special case of 'Cronbach's alpha' that applies to binary items) estimate of reliability that can be written as

$$\alpha = (k/(k-1))(1 - (\sum_{j=1}^{k} P_j(1-P_j))/\sigma_x^2) \quad (5)$$

where $P_j$ is the proportion of the sample with correct answers to item *j.* We note that there is no inherent assumption of 'unidimensionality' necessary here. The underlying concept is one where the ratio of true to observed score is conditioned on the set of sampled individuals and correlated item sets. As the number of items increases so the reliability estimate will tend to 1.0. However, the conditional independence assumption is crucial. We can see this by considering the two half-test total scores as $d_1, d_2$ where we have

$$var(d_{i1} - d_{i2}) = var(d_{i1}) + var(d_{i2}) - 2cov(d_{i1}, d_{i2}). \quad (6)$$

The last term in (6) is zero conditional on individual true ability independence, but if this conditional independence assumption is violated, for example if the covariance is positive, as might often be the case if a (perceived) correct answer to an item increases the probability of a correct response to following items, the estimator in (5) will tend to overestimate the reliability. In effect, internal consistency estimates are attempting to estimate the sampling variance associated with the sum (or mean) of *k* item responses where the probability of a positive item response is determined by a, possibly complex, function of item 'difficulty' and each individual's 'true' ability. In the simulation reported below we provide one example of such an underlying model which allows for positive dependency among item responses and shows that ignoring this leads to an overestimate of the reliability using internal consistency estimates. One particular problem with this approach is that there is typically no way to validate the conditional independence assumption. This leads us to consider an alternative approach based on the use of instrumental variables.

## 3.2 Instrumental variable (IV) estimation

Consider first the case of a simple regression model as in (2). Suppose we have a variable *Z* where we assume a linear model relating $x$ to $Z$, namely

$$x_i = \gamma_0 + \gamma_1 Z_i + e_z, \quad \gamma_1 \neq 0, \tag{7}$$

where *Z* is uncorrelated with the random terms in (2). We shall return to this key assumption below. A two-stage procedure can be used, where the predicted values $\widehat{x_i}$ from (7) replace the explanatory variable, $X_i$, in the model of interest (2) to obtain consistent estimates for the parameters of that model. It can be shown that the parameter $\beta_1$ is estimated by $(Z^T x)^{-1} Z^T y$ with estimated covariance matrix $cov(\widehat{\beta_1}) \cong \sigma_z^2 (Z^T x)^{-1} (Z^T Z)(x^T Z)^{-1}$ where $\sigma_z^2$ is estimated from the empirical residuals in the regression of $y$ *on* $\hat{x}$. The estimate for the reliability is then obtained as $R_x = \alpha_1/\beta_1$. Where we have several explanatory variables measured with error and where the measurement errors may be correlated, $X$ and $Z$ are now vectors and without loss of generality we assume that *Z* contains both variables with measurement error and those without, for whom the measurement error variance is zero. The estimators then have the same form as those given above. We also have the usual estimator from (3), $cov(\hat{\alpha}_1) = \sigma_{e_x}^2 (x^T x)^{-1}$, which allows us to obtain consistent estimators for the variances and covariances of the measurement errors via differencing, and hence we can form the measurement error matrix. A detailed description of such IV estimators can be found in Carroll et al. (2006, Chapter 6), who also provide a detailed description of estimation methods where measurement error exists.

We can also consider joint IV model estimation rather than the two-stage procedure, although the latter will generally be satisfactory for large samples. Write

$$\begin{aligned} Y_i &= \beta_0 + \beta_1(\gamma_0 + \gamma_1 Z_i) + e_X \\ x_i &= \gamma_0 + \gamma_1 Z_i + e_x, \quad \gamma_1 \neq 0, \\ Z_i &= \delta_0 + e_Z \\ E(e_X e_x) &\neq 0,\ E(e_x e_Z) = E(e_X e_Z) = 0. \end{aligned} \tag{8}$$

The likelihood is then proportional to the separate likelihood contributions from these three sub-models and we can fit, for example, a Bayesian model with suitable, say diffuse, priors using MCMC to update the parameters, in particular $\beta_1$.

It may be the case, as suggested above, that the measurement error variance depends on further factors, assumed to be measured without error, for example it may differ for males and females. In simple cases a separate analysis for each group may be satisfactory, but in general we may wish to model the dependency. In this case these variables can be added to the first two lines of (8) or in the equivalent 2-stage procedure, with suitable interaction terms where necessary, that allow $\beta_1$ to vary as a function of these factors. These varying measurement errors can be incorporated for measurement error adjustment within the final model of interest, even if they do not explicitly appear in that model.

As a simple example suppose we have sex $(S)$ coded (0: male;1: female). In our 2-stage procedure we first estimate the prediction model

$$x_i = \gamma_0 + \gamma_1 Z_i + \gamma_2 S_i + \gamma_3 Z_i S_i + e_x$$
$$Y_i = \beta_0 + \beta_1(\gamma_0 + \gamma_1 Z_i + \gamma_2 S_i + \gamma_3 Z_i S_i) + e_X \qquad (9)$$

which is then compared with the model for the observed predictor

$$Y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 S_i + \alpha_3 S_i x_i + e_X^*$$

for each value of $S$. In the simple case this provides separate estimates for each sex. The assumption that the IV *Z* is uncorrelated with the random residual terms in the first two lines of (9) needs to be discussed and justified in practice. In the case we consider in this paper, where a distal variable is used as the IV, we can typically appeal to the existence of a relatively long time gap to ensure that *Z* is uncorrelated with *m* in (1). We now consider the conditions where the second assumption $E(Ze_X) = 0$, will be satisfied, at least approximately.

## 3.3 Assumptions for IV estimation

We consider for simplicity the multivariate normal case

$$\begin{pmatrix} y \\ X \\ Z \end{pmatrix} \sim N(0, \Omega), \quad \Omega = \begin{pmatrix} \sigma_1^2 & & \\ c_{12} & \sigma_2^2 & \\ c_{13} & c_{23} & \sigma_3^2 \end{pmatrix}. \qquad (10)$$

We allow for the possibility that the IV variable *Z* may or may not have its own independent measurement errors and this is also the case for $y$. To satisfy the assumption $E(Ze_X) = 0$ we require

$$E\big(Z(y - c_{23}X)\big) = \sigma_2^2 c_{13} - c_{23}c_{12} = 0$$

or equivalently in terms of correlations $\rho_{13} - \rho_{23}\rho_{12} = 0.$ (11)

We shall return to a discussion of this assumption in our simulation and example. The assumption that *Z* is uncorrelated with the measurement errors, in the case of longitudinal data, will typically be satisfied by a suitable choice of distal measure, implying that any direct path from the observed distal measure to *Z* operates solely through the true value of the distal measure. We shall return to this issue.

## 3.3 IV grouping estimators

A commonly advocated, but typically unsatisfactory, IV method is the so-called grouping procedure, first suggested by Wald (1940). This is based upon dividing the observed data into groups but is typically put forward without reference to certain basic assumptions that are required. Since it is sometimes advocated (see below), in this section we briefly explain the procedure and demonstrate, in a straightforward fashion, the problems associated with its use in practice.

The standard measurement error model is written as in (1)

$$x_i = X_i + m_i\ ,\ y_i = Y_i$$

where capital letters refer to the true values, and we assume no measurement error in the response. The model of interest is, as before

$$E(Y_i|X) = \beta_0 + \beta_1 X_i \quad (12)$$

whereas what we observe is

$$E(Y_i|x_i) = \alpha_0 + \alpha_1 x_i. \quad (13)$$

The proposed Wald (1940) instrumental variable estimator for the true regression slope $\beta_1$ can be written as

$$[E(y_i|x_i > \mu) - E(y_i|x_i \leq \mu)]/[E(x_i|x_i > \mu) - E(x_i|x_i \leq \mu)] \quad (14)$$

where expectations are replaced by observed means and $\mu$ is taken as the median of $x_i$. For a simple regression model given by

$E(X_{2i}) = \beta_0 + \beta_1 X_{1i}$

it follows that, for the *n* observations in the interval [a,b] and replacing $X_1$ with the observed means, then we can write

$$(\frac{1}{n}) \sum_i X_{2i} = (\frac{1}{n})(n\beta_0 + \beta_1 \sum_i X_{1i}) = \beta_0 + \beta_1 (\frac{1}{n}) \sum_i X_{1i}. \quad (15)$$

In other words, over the interval, the point defined by the means of the response and explanatory variables lies on the regression line. This will thus be the case for both models (12) and (13). Hence, for the estimator implied by (14) where the groups are defined by the median of the observed *x*, the respective means below and above the median both lie on the line defined by the observed regression (13) so that (14) in fact estimates the observed regression slope and not the true one. This will generally be true for all grouping estimators, including methods that use weighted functions of (*x*,*Y*) where the weights are defined using *x* rather than X (e.g. Durbin, 1953). The problem is that the conditioning is of necessity based upon the observed values of the explanatory variable rather than the unknown true values.

Wald (1940, p.294-295) distinguishes two rules. The first groups the sample on the observed *x* values around the median (or some other value). The second rule considers the case where the sample can be grouped on the basis of the true values. He points out that the first rule is invalid since the grouping is not independent of the measurement errors and then assumes that the measurement error itself (*m*) is bounded by an interval $[-c, c]$ and that all the values $x_i, i = 1, \ldots, N$, lie outside the interval

$$[\mu - c, \mu + c] \ . \quad (16)$$

Clearly in this case those observed to be above the median (based on the observed data but in expectation equal to the median of the true data) are the same set as those true values above the median and likewise for those below the median. Wald (1940) then shows that if the probability that *m* lies outside the above interval is negligible (as is the proportion of observed *x* lying inside the interval $[\mu - c, \mu + c]$ ), then clearly the means defined according to his rule 1, on which (12) is based, will be good approximations to the true means and hence gives us a consistent estimator of the true slope. A similar set of assumptions for consistency is also required for the procedures suggested by Bartlett (1949)

and Durbin (1954). Neyman and Scott (1951) derive a similar, although more general, result for grouping estimators.

The major problem is that condition (16) will only hold for certain distributions, typically where the density around the median is a minimum. Otherwise, for example for unimodal symmetric distributions such as the normal, and also for unimodal skew distributions, the value of *c* would need to be very small so that the measurement error variability likewise would need to be very small. Thus, where the true values follow a standard normal distribution with a sample size of just 1000 and using a measurement error of just 0.05 implying a reliability of about 0.95, the mean absolute value of the differences between successive observed values is about 0.006, whereas the mean absolute value of the measurement errors is about 0.2 which is actually, with a very high probability, greater than any difference between successive observed values. In other words, for assumption (16) to hold to a good approximation, we would require such a small value of *c* that we could anyway effectively ignore measurement errors.

In the standard econometrics literature that quotes these grouping IV methods, (see for example Johnston, 1972, Cameron and Trivedi, 2005), one does not, unfortunately, find reference to condition (16), despite the fact that it does appear to be crucial. Fuller (2006, chap 1.6) does mention it, but just in passing.

## 4. Using distal test scores as IVs

In Section 5 we discuss an example using longitudinal education achievement data from Queensland, Australia. This uses a distal test score as an instrumental variable, namely a measure of attainment at year 3 of schooling when estimating the reliability of a year 5 attainment score. We tend to obtain estimates of *R* that are approximately 10% lower than those obtained using internal consistency estimates. One potential inference from this is that the assumption upon which the internal consistency measures are based, that of local independence, may be invalid since, as we have shown in Section 3.1, a positive dependency between items biases upwards internal consistency measures, and in a separate paper we look at ways of estimating a parameter for the dependency. This lack of independence will also lead to biases in the case of measures based upon latent variable models. One of these

is the Omega coefficient (McDonald 1999) that posits a unidimensional factor model for the set of items, fixes the factor variance at 1.0 and then derives the reliability from the set of factor loadings. Item response models such as the Rasch model adopt a similar approach and in all these cases local independence is assumed. It is this independence assumption rather than the unidimensionality assumption that is important. If the local independence assumption fails, then these coefficients like Cronbach's alpha, will not produce consistent estimates of reliability. Moreover, the estimate of the true score variance will depend crucially on the assumed model form, even if we do have independence. In the next section we shall demonstrate the effect of dependency among items using a simulation for a set of binary item responses.

## 4.1 A simulation for distal score IV estimation

We demonstrate the use of distal scores through a simulation where we assume we have data on 10,000 individuals measured at three occasions (for example three successive years of primary schooling) with correlated true scores across occasions and a sample of items chosen from a distribution of test items to produce a set of binary response items with increasing amounts of dependency. We fit the internal consistency estimates and compare with the true reliabilities as determined by the data generating mechanism and with the distal IV estimates.

Each test score is based on a $k$ = 30 item test. For convenience we assume a latent probit model where the probability of observing a correct response $y_{ij,t} = 1$ for individual $i$ to item $j$ at occasion $t$ =(1, 2, 3) is modelled as follows.

We have covarying true scores $\theta_{i,1}, \theta_{i,2}, \theta_{i,3}$ across three occasions distributed as

$$\begin{pmatrix}\theta_{i,1}\\ \theta_{i,2}\\ \theta_{i,3}\end{pmatrix} \sim N\left\{\begin{pmatrix}0\\0\\0\end{pmatrix}, \begin{pmatrix}0.25 & & \\ q & 0.25 & \\ 4q^2 & q & 0.25\end{pmatrix}\right\} \qquad (18)$$

where the one-occasion apart covariance $q$ will take the values (0.1, 0.125, 0.150, 0.175) corresponding to one-occasion apart correlations $\rho_{12} = \rho_{23}$ as (0.4, 0.5, 0.6, 0.7), and two-occasion apart correlations $\rho_{13}$ (0.16, 0.25, 0.36, 0.49). We note that this covariance pattern will generate consistent IV estimates satisfying (11), and because the Bernoulli responses

are sampled independently *across* occasions the remaining assumption in (8) for consistency of IV remains valid.

The probability of observing a correct response $y_{ij,t} = 1$ for item *j* at a given occasion, is defined as

$$p_{ij,t} = \int_{-\infty}^{\theta_{i,t}-\alpha_{j,t}} \phi(z)dz, \qquad (19)$$

where $\phi(z)$ is the standard normal distribution, and we sample the response $y_{ij,t}$ as (0,1) from a Bernoulli distribution with this probability. The parameters $\theta_{i,t}, \alpha_{j,t}$ represent individual ability and item difficulty respectively. We use this simple model for illustration purposes only, but our general results will apply for more complex models. Model (19) is essentially the probit version of the common '1 parameter' (logistic) item response model (Rasch model). Thus for each individual we obtain 30 binary responses and the sum of these responses is the observed score. The observed score for individual *i* is defined as

$$y_{i,t} = \sum_{j=1}^{30} y_{ij,t}$$

and the observed score variance for each occasion is then estimated from the observed scores simulated for 10,000 students.

We assume that basic measurement error is introduced as a result of the item sampling, and there is no further measurement error. For each individual and occasion, we therefore simulate *N* draws for the items, where we choose *N*=250, and obtain, at occasion *t*, *N* values of $p_{ij,t}$. From these $p_{ij,t}$ we sample from the Bernoulli distribution to obtain a (0,1) response, say $y_{ij,t}$, for an item and compute the test score $\sum_{j=1}^{k} y_{ij,t}$ by summing the binary responses for the items. The average of these scores over the *N* draws is taken as the individual's true value from which we obtain the between-individual true score variance at each occasion. For each individual and occasion, we compute the between-draw variance for the test score $\sum_{j=1}^{k} y_{ij,t}$ and average these over individuals to obtain an estimate of the measurement error variance. An estimate of the reliability is then computed using the simulated true score variance and the simulated observed score variance. Also, for each draw we compute the internal consistency estimate based upon the $y_{ij,t}$ and average these across draws. The IV estimates of reliability are also computed.

We assume that test candidates work through the test items in order. To induce dependency across item responses we proceed as follows. Starting with item 1 we note whether the outcome is a success or not. If the former, then to sample the second item response we modify $\theta_{it} - \alpha_{jt}$ by adding a chosen value *c* where, of course, c = 0 implies local independence. If the previous response is a failure, then we subtract this value *c*. We note that, unlike a standard autoregressive formulation, this dependency respects the item ordering and is not symmetrical. We then sample the response to item 2 and repeat the process, until we have sampled all items. The simulations were implemented using Stata software, version 15 (StataCorp, 2017). The model (19) for item 1 thus becomes

$E(y_{i1}) = prob(y_{i1} = 1) = p_{i1}(\theta_i, \alpha_1) = \int_{-\infty}^{\theta_i - \alpha_1} \phi(z)dz$ , $\phi \sim N(0,1)$

and for $j > 1$

$$p_{ij}(\theta_i, \alpha_j) = E(y_{i,j-1})(\int_{-\infty}^{\theta_i - \alpha_j + c} \phi(z)dz) + [1 - E(y_{i,j-1})](\int_{-\infty}^{\theta_i - \alpha_j - c} \phi(z)dz)). \quad (20)$$

In Table 1 we compare the reliability estimates based on the true score variance and the observed score variance as generated from the simulation. Our principal interest lies in understanding how each procedure operates under different amounts of inter-item dependency, represented by increasing values of the parameter *c*=(0, 0.25, 0.50, 0.75, 1.0).

Table 1 shows that in the case of independence among items (c=0) the true reliability, based on the simulated true scores relative to the simulated observed scores, is well approximated by the internal consistency estimate of reliability.

(Table 1 here)

Increasing the value of *c* leads to overestimation of the true reliability by the internal consistency estimate whereas the IV estimator of reliability R(IV) decreases with $c$, and closely approximates the true reliability that also decreases with $c$. As is evident from this Table, upward biases of some 20% can be obtained from use of coefficient alpha, with corresponding biases for parameter estimates in models based on such alpha estimates. Here, $\rho_{12}$ is the one-occasion apart Pearson correlation between the true scores implied by $q$. Intuitively, we can envisage that as *c* increases, the underlying value $\theta_i + c - \alpha_j$ and $\theta_i -$

$c - \alpha_j$ generating the set of observed (0,1) responses are dominated by *c* so that the variation due to the true parameters $\theta_i$ decreases and hence the (true) reliability decreases. We note that even relatively small values of *c* (on the standard normal probability scale) are associated with marked changes in the reliability. We note that (19) and (20) are here used to explore the behaviour of the reliability as the adjacent item association changes. It is in fact a simple case of a novel class of item response models that has received little if any attention in the psychometric literature. We shall return to this model in the discussion, noting again that here, for our purposes, it is introduced simply to demonstrate the effect of a violation of the conditional independence assumption.

### 4.1.1 Simulating IV estimates

We now consider a simulation for IV estimates that allows the assumption given by (11) to be violated. We assume that the IV is uncorrelated with measurement errors at another occasion, as discussed in Section 3.2. We have simulated normally distributed observed scores at 3 occasions $t = 1,2,3$ each with zero mean and variance 1.0. The reliability of the observed scores at occasion 1 and 3 is set equal to $0.8$. The reliability of the observed scores at occasion 2, $R_2$, is allowed to vary across the different conditions of the simulation study. The first order correlations between the true scores $\rho_{12}$ and $\rho_{23}$ are set equal to 0.5. The assumption (11) will hold when the second order correlation $\rho_{13}$ is equal to 0.25, the product of the first order correlations. We shall vary this correlation across the different conditions of the simulation study. We conduct seven simulation studies. In studies 1, 2 and 3 we set $R_2$ to be 0.7, 0.8 and 0.9 respectively. In all three cases we set $\rho_{13} = 0.25$ and so in these studies (11) holds. In studies 4, 5, 6, and 7 we set $R_2 = 0.8$, but we vary $\rho_{13}$ to be 0.15, 0.20, 0.30, and 0.35, respectively. Thus, in these four studies the assumption (11) fails to differing degrees. In each study we conduct 1000 replications.

(Table 2 here)

Table 2 shows that the $50\text{th}(\hat{R}_2)$ percentile, the median estimate of $R_2$, is unbiased in studies 1-3 where we set the correlation between the first and third and occasion true scores $\rho_{13}$ to be 0.25 (and therefore (11) holds), but where we vary the reliability of the second occasion observed scores $R_2$. However, the $50\text{th}(\hat{R}_2)$ percentile is biased upwards in studies 4 and 5 where we lower the correlation between the first and third and occasion

true scores $\rho_{13}$ from 0.25 to 0.15 and 0.20, and is biased downwards in studies 6 and 7 where we raise the correlation between the first and third and occasion true scores $\rho_{13}$ from 0.25 to 0.30 and 0.35. It illustrates the importance for the correlation assumptions of the proposed IV estimate.

## 5. An example using NAPLAN data for Queensland

Since 2008, the Australian National Assessment Program–Literacy and Numeracy (NAPLAN) has been used to report progress in student achievement and to compare the performances of different groups and is a major focus of Australian education policy. In this example we use Numeracy data for the Queensland cohort of pupils who participated in NAPLAN in Year 3 in 2011, Year 5 in 2013 and Year 7 in 2015, with approximately 53,000 students from 1,400 government and non-government elementary schools. See Cumming, Goldstein and Hand (2019) for more details. In the present analyses we have only used individuals with scores present at all occasions in our models which reduces the sample size to approximately 50,000. We have independently normalised each of the scale scores.

We see from the simulation results that if the local independence assumption fails and item responses have a positive dependency (e.g students have a run of successes), then internal consistency estimators such as KR20 may overestimate reliability. In general we might expect the conditional independence assumption to be false for a number of reasons, not least that a student with a given 'true' attainment who happens, at a particular session, to answer an item incorrectly will often be aware of this, especially when items are ordered by difficulty, and this is likely to influence their propensity to correctly answer subsequent items. For the IV method, since we are using a year 3 score as the IV, it seems reasonable to assume that it is uncorrelated with the measurement error in the year 5 score. We also would argue that it is likely to have a negligible correlation with the residual in the model of interest which uses the year 7 score as a response with the year 5 score as a covariate, so that (11) will be satisfied. This residual measures the relative progress made by a student given their attainment at year 5. We also carry out some sensitivity analysis around this assumption.

We estimate the reliabilities as follows. In the first stage, for each attainment test we first regress the year 5 score on the year 3 corresponding score. This provides predicted values for the year 5 scores and these are then used as the covariate in a regression with year 7 score as outcome. The year 7 score is also regressed on the observed year 5 score and the reliability is obtained from the ratio of the regression coefficients as described in section 3.2. We have also tried adding further predictors for the year 5 score and fitting simple 2-level random-intercept multilevel models with the year 5 school identifying the second level, but the estimates hardly change. Additionally, we have studied reliability by student characteristics such as gender and indigenous status, and obtain very similar results.

(Table 3 here)

Table 3 shows the results for both distal IV estimates and internal consistency estimates for a set of numeracy test scores.

The estimates for the IV procedure are all lower by between 4% and 10% of the internal consistency estimates consistent with the argument that the alpha coefficients are biased upwards in the presence of positive item dependence, given that some dependency is to be expected. We also note that the IV estimates do not depend on any assumptions about the internal test score structure such as local independence, since they operate at the level of the test score. To illustrate the importance of independence among measurement errors, in Table 4 we compare the use of a contemporaneous measurement at year 5, the numeracy total, as the IV for the four subtests where by definition the subtest measurement errors are a component of the measurement error for the total score, so violating the conditional independence assumption.

(Table 4 here)

We see that the use of the same year total maths score leads to overestimates of reliability relative to the internal consistency estimates, as well as the IV estimates where Year 3 score is used as the instrument; up to about 20%. If we use a different subtest score as IV that is measured at the same time as the target variable we also find that the estimate of reliability tends to increase. In other words it appears that the observed IV test score contains less measurement error than using a distal test score so that we are not fully correcting for it (assuming the distal test score gives consistency), because the test score at the same time shares measurement error with the IV measure and the score that is predicted from it will also contain the shared measurement error.

As expected, the reliability estimates also increase with the number of test items. Using as additional IV variables, measures from different domains to predict year 5 scores, gives results that are very similar. Nevertheless, in practice it is recommended that the sensitivity of the IV estimates is explored using different combinations of IV predictors. To illustrate we have used different combinations of IV variables for the year 5 measurement test score estimate where we have also used available year 9 scores. The results are given in Table 5.

(Table 5 here)

We see from Table 5 that all three prediction models that use the year 3 measurement give similar results, suggesting that a value that is the mean of these, 0.64, would be a suitable value, with perhaps a sensitivity analysis using the minimum and maximum of all the estimates, which in this case should include the internal consistency estimate.

## 5.1 Fitting the model of interest

Table 6 shows the changes to the model of interest for a simple 2-level random-intercept multilevel model (students within schools) where the measurement score at year 7 is the response and the corresponding score at year 5, together with gender and indigenous status are covariates, both being binary variables. We have fitted our measurement error model (1)-(3) assuming four different reliability values, from completely reliable ($R$=1.0) to a low value ($R$=0.58).

(Table 6 here)

We see that for the range of values of $R$ chosen, there is a far smaller effect associated with Indigenous status than in the unadjusted analysis ($R$=1). In particular, as the value of $R$ decreases, so the lower progress made by indigenous students is more confined to those students with higher year 5 scores. A similar finding holds for female students. Over the range of values of $R$ (<1) it appears that with decreasing values of $R$ the apparent effect due to Indigenous status gets smaller. This clearly is of considerable relevance educationally and illustrates the importance of paying attention to obtaining good estimates of reliability and measurement error variances. The value of $R$ given by coefficient alpha is close to the median value suggested from Table 6, and this may be somewhat reassuring for those studies that have relied upon such an estimate, but not a good reason for not exploring alternatives such as IV procedures.

## 6. Discussion

We have illustrated, that with suitable longitudinal data, instrumental variable estimates of test score reliability can provide reasonable values for reliabilities that can then be used in further models to adjust for test score measurement error. While our focus has been on scaled scores, illustrated using binary component items, the IV estimates we propose are generally applicable to scores however derived.

In the case of administrative data collected for accountability purposes, such as the NAPLAN data in Australia, there will generally be a lack of cross sectional variables suitable for use as instruments, but there will typically be longitudinal data that can be utilised to estimate measurement error. Both IV estimators and internal consistency estimators make particular, but different, assumptions about the underlying relationships. The analyst can adopt a conservative procedure as suggested in the preceding section by conducting a range of sensitivity analysis over the range of estimates, but we may also be able to study the plausibility of the different assumptions, using data that are available.

In the case of longitudinal distal IV estimation, we have the problem of choosing a suitable IV variable. One possible procedure would be to construct or choose a set of scale scores at each longitudinal occasion with increasing numbers of items to reflect increasing reliability values. We would then examine the expression given in (11) (using the actual observed values for the correlations) and utilise these to extrapolate to a test with a high enough number of items to approximate a reliability of 1.0 and thus to infer whether condition (11) holds, at least approximately. Where standardised tests or scale scores are used in research studies, careful attention should be paid, by test and scale constructors, to providing plausible estimates of measurement error distribution parameters, and the use of IV estimators in addition to existing methods, can be helpful. When results from analyses that use such measures are reported, it should also become standard practice to discuss the role of measurement errors.

Where suitable IV variables exist within the dataset being analysed, then these can be utilised directly as in model (8). Because the existence of such variables is likely to be uncommon, a more general approach is to provide analysts with suitable estimates for

measurement error or reliability derived for the measurements being used, by test constructors.

The IV estimators are useful where conditional independence assumptions for scaled scores are likely to be violated, but also in the common case where item data to compute internal consistency measures are unavailable or where there is some doubt about the validity of any measures that may be supplied by, for example, test constructors. Our simulations demonstrate that when the assumption of conditional independence is violated and there is positive dependency between items standard internal consistency estimators such as Cronbach's alpha are likely to be biased upwards, although in our practical example the difference seems no larger than about 10%. Nevertheless, it would seem prudent to require providers of external reliability estimates based upon coefficient alpha or similar procedures, to justify that assumption, since there are clearly situations when it would be expected to fail such as when the probability of a correct response to an item is deliberately designed to rely upon a correct response to a previous item.

In our exposition and simulation we have considered a standard case where an individual responds to a pre-existing common set of test items. In other cases, such as computer adaptive testing (CAT) the next item in a test for an individual will depend on their previous responses. Thus the 'c' parameter in our simulation will itself be a function of previous responses, for example reflecting positive correlations between items followed by a negative one. Although our IV procedure does not depend explicitly on parameters such as 'c', since it operates at the total score level, it would be an interesting piece of further research to ascertain the effect on, for example, the estimate of coefficient alpha.

The estimation algorithm is that described by Goldstein et al. (2017) and is available from the first author. It is implemented in MATLAB (Mathworks, 2019). The code used for generating the simulated data is written in Stata (Statacorp, 2017) and is available from the second author.

Finally, we note that in our simulation we have introduced what appears to be a novel elaboration of the standard probit one parameter item response model (Rasch model) which allows for a simple form if item dependency. Further work on fitting such a model and more complex unidimensional and multidimensional models is planned, especially with a

view to exploring the extent to which such models may be useful in the estimation of reliability.

## Acknowledgements

This work received part support by a grant from the Economic and Social Research Council (UK) award number ES/R010285/1 and part support from the Institute for Learning Sciences and Teacher Education, Australian Catholic University. The example in this research was applied to data from the Queensland Department of Education. The views expressed in this paper are those of the authors alone and do not represent those of the Department. The authors take responsibility for the integrity and accuracy of the analysis. There are no conflicts of interest

## References

Bartlett, MS. (1949). Fitting a straight line when both variables are subject to error. *Biometrics*, 5, 207-212.

Cameron, CA, and Trivedi, PK (2005). *Microeconometrics*. Cambridge, New York.

Carroll, R., Kuchenoff, H., Lombard, F., and Stefanski, LA. (1996). Asymptotics for the SIMEX Estimator in Nonlinear Measurement Error Models. *J. American Statistical Association*, Vol. 91, No. 433 (Mar., 1996), pp.242-250.

Carroll, RJ, Ruppert, D, Stephanski, LA and Crainicaenu (2006). *Measurement error in nonlinear models*. Boca Raton, Chapman and Hall.

Cook, JR., and Stefanski, LA. (1994). Simulation-Extrapolation Estimation in Parametric Measurement Error Models. *J. American Statistical Association*, Vol. 89, No. 428 (Dec., 1994), pp.1314-1328

Cumming, J., Goldstein, H, and Hand, K. (2019). Use of Educational Data to Measure Educational Progress: Australian Accountability Policy, Data and Indigenous Student Outcomes. (submitted for publication).

Durbin, J. (1954). Errors in variables. *Rev. International Statistical Institute*, 22, 23-32.

Ecob R & Goldstein H. (1983). Instrumental Variable Methods for the Estimation of Test Score Reliability. *Journal of Education Statistics*. 8 (3) 223-241.

Frederic M. Lord, Melvin R. Novick (2008). *Statistical theories of mental test scores* (reprinted). Information Age Publishing, Charlotte, NC.

Fuller, WA (2006). *Measurement error models*, New York, Wiley.

Goldstein, H., Browne, WJ., and Charlton, C. (2017). A Bayesian model for measurement and misclassification errors alongside missing data, with an application to higher education participation in Australia. *Journal of Applied Statistics*. doi: 0.1080/02664763.2017.1322558

Goldstein, H., Kounali, D., and Robinson, A. (2008). Modelling measurement errors and category misclassifications in multilevel models. Statistical Modelling; 8(3): 243-261.

Johnston, J. (1972). *Econometric Methods*, second edition. McGraw Hill, New York.
Mathworks (2019). Matlab. https://uk.mathworks.com.

Leckie, G. and Goldstein, H. (2019) The importance of adjusting for pupil background in school value-added models: A study of Progress 8 and school accountability in England. *British Educational Research Journal*, 45, 518-537.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum.

Meier, E., Spierdijk, L., and Wansbeek, T. (2017). Consistent estimation of linear panel data models with measurement error. Journal of Econometrics 200 (2017) 169–180

Neyman, J. and Scott, EL. (1951). On certain methods of estimating the linear structural relation. *Ann. Math. Statist*., 22, 352-361.

StataCorp. 2017. *Stata Statistical Software: Release 15*. College Station, TX: StataCorp LLC

Wald, A. (1940). The fitting of straight lines when both variables are subject to error. *Ann. Maths. Statist.*, 11, 284-300.

## TABLES

Table 1. Reliability estimates for 30 items from 3-occasion longitudinal data: means for simulated 250 draws from each of 10,000 iterations, for selected values of q and c. Also given in brackets are standard errors (SE) for the IV estimates. The rows labelled 'R(true)' refer to the value of R derived directly from the simulated true scores and the simulated observed scores. The rows labelled 'R(alpha)' are those estimates obtained using coefficient alpha estimates (internal consistency). The rows labelled 'R(IV)' are those obtained from the proposed instrumental variable estimator.

| | c=0 | c=0.25 | c=0.50 | c=0.75 | c=1.0 |
|---|---|---|---|---|---|
| q=0.1, $\rho_{12}$=0.40 | | | | | |
| R (true) | 0.815 | 0.809 | 0.791 | 0.762 | 0.727 |
| R(alpha) | 0.815 | 0.868 | 0.901 | 0.926 | 0.945 |
| R(IV) | 0.858 | 0.765 | 0.777 | 0.717 | 0.730 |
| R(IV) SE | (0.042) | (0.036) | (0.039) | (0.044) | (0.049) |
| q=0.125, $\rho_{12}$=0.50 | | | | | |
| R (true) | 0.818 | 0.807 | 0.793 | 0.765 | 0.729 |
| R(alpha) | 0.818 | 0.867 | 0.903 | 0.926 | 0.945 |
| R(IV) | 0.800 | 0.777 | 0.796 | 0.763 | 0.707 |
| R(IV) SE | (0.023) | (0.022) | (0.025) | (0.025) | (0.028) |
| q=0.15, $\rho_{12}$=0.60 | | | | | |
| R (true) | 0.811 | 0.805 | 0.790 | 0.761 | 0.731 |
| R(alpha) | 0.811 | 0.865 | 0.901 | 0.925 | 0.946 |
| R(IV) | 0.812 | 0.809 | 0.801 | 0.769 | 0.739 |
| R(IV) SE | (0.016) | (0.018) | (0.018) | (0.019) | (0.022) |
| q=0.175, $\rho_{12}$=0.70 | | | | | |
| R (true) | 0.815 | 0.807 | 0.791 | 0.765 | 0.729 |
| R(alpha) | 0.815 | 0.866 | 0.902 | 0.927 | 0.945 |
| R(IV) | 0.818 | 0.794 | 0.778 | 0.748 | 0.730 |
| R(IV) SE | (0.011) | (0.011) | (0.013) | (0.013) | (0.017) |

**Table 2. Reliability estimates obtained using IV with simulated data. 1,000 replications were used. Quantile estimates for the reliability of the second occasion measurement are shown for different sample sizes and occasion 2 reliabilities $R_2$ and correlation between occasion 1 & 3 true scores $\rho_{13}$.**

| Study | N | $R_2$ | $\rho_{13}$ | $\sigma_{13}$ | 50th($\hat{R}_2$) | 2.5th($\hat{R}_2$) | 97.5th($\hat{R}_2$ |
|---|---|---|---|---|---|---|---|
| 1 | 1000 | 0.70 | 0.25 | 0.20 | 0.695 | 0.532 | 0.950 |
| 2 | 1000 | 0.80 | 0.25 | 0.20 | 0.803 | 0.631 | 1.067 |
| 3 | 1000 | 0.90 | 0.25 | 0.20 | 0.900 | 0.707 | 1.205 |
| 4 | 1000 | 0.80 | 0.15 | 0.12 | 1.320 | 0.931 | 2.303 |
| 5 | 1000 | 0.80 | 0.20 | 0.16 | 0.996 | 0.751 | 1.486 |
| 6 | 1000 | 0.80 | 0.30 | 0.24 | 0.664 | 0.528 | 0.841 |
| 7 | 1000 | 0.80 | 0.35 | 0.28 | 0.572 | 0.458 | 0.709 |

**Table 3. Reliability estimates comparing Cronbach's alpha with observed year 5 patterns and IV methods based upon year 3 scores as instrumental variable in the regression of year 7 on year 5 scores.**

| | Algebra, Function & Pattern (4 items)* | Measurement, Chance and Data (13 items) | Number (13 items) | Space (10 items) | Numeracy Total (40 items) |
|---|---|---|---|---|---|
| Year 5 IV | 0.403 | 0.648 | 0.625 | 0.528 | 0.793 (0.790)** |
| Year 5 coefficient alpha | 0.449 | 0.678 | 0.681 | 0.576 | 0.864 |

* No year 3 algebra so year 5 regressed on Total numeracy at year 3. **The term in brackets is based upon using a 2-level model.

| **Table 4. Reliability estimates year 5 subtest scores with IV as year 5 total numeracy score.** | | | | |
|---|---|---|---|---|
| | Algebra, Function & Pattern (4 items) | Measurement, Chance and Data (13 items) | Number (13 items) | Space (10 items) |
| Year 5 IV | 0.512 | 0.799 | 0.647 | 0.673 |

| **Table 5. Reliability estimates for measurement test score at 5 years using different combinations of IV variables. All models are additive linear regression models.** | |
|---|---|
| **IV variables predicting year 5 measure score** | **Reliability estimate** |
| Year 3 measurement chance and data | 0.648 |
| Year 9 measurement chance and data | 0.585 |
| Year 3 total numeracy | 0.722 |
| Year 3 total numeracy + year 3 measurement chance and data | 0.645 |
| Year 3 total numeracy + year 9 measurement chance and data * | 0.616 |
| Year 3 measurement + year 9 measurement chance and data | 0.655 |
| * A model that also included the year 3 measurement chance and data was fitted but the latter coefficient was small and not significant at 5%. | |

**Table 6. Measurement score at year 7 related to year 5 measurement score, gender and indigenous status. Differing reliability values for year 5 score. Year 7 reliability set to 0.75. Standard errors in brackets. Burn in = 100, iterations = 250.**

| **covariate** | ***R*=1** | ***R*=0.72** | ***R*=0.64** | ***R*=0.58** |
|---|---|---|---|---|
| Intercept | 0.009 (0.008) | -0.039 (0.008) | -0.058 (0.009) | -0.077 (0.007) |
| Year 5 score | 0.658 (0.005) | 0.947 (0.007) | 1.067 (0.008) | 1.170 (0.015) |
| Female | -0.038 (0.007) | 0.022 (0.007) | 0.046 (0.009) | 0.067 (0.007) |
| Indigenous | -0.283 (0.020) | -0.072 (0.022) | 0.017 (0.021) | 0.099 (0.021) |
| Year 5 x female | -0.059 (0.007) | -0.086 (0.009) | -0.088 (0.009) | -0.086 (0.009) |
| Year 5 x Indigenous. | -0.088 (0.015) | -0.088 (0.019) | -0.077 (0.019) | -0.065 (0.019) |
| Female x indigenous | 0.000 (0.003) | -0.024 (0.027) | -0.023 (0.029) | -0.021 (0.026) |
| $\sigma_e^2$ | 0.386 (0.002) | 0.268 (0.003) | 0.217 (0.002) | 0.170 (0.006) |
| $\sigma_u^2$ | 0.030 (0.002) | 0.030 (0.002) | 0.030 (0.002) | 0.030 (0.002) |
| $\sigma_e^2$ is between student variance, $\sigma_u^2$ is between school variance. | | | | |