CENTRE FOR
SOCIAL SCIENCE RESEARCH

Southern Africa Labour and Development
Research Unit

TESTING FOR A COMMON LATENT VARIABLE IN A LINEAR REGRESSION: OR HOW TO "FIX" A BAD VARIABLE BY ADDING MULTIPLE PROXIES FOR IT

Martin Wittenberg

CSSR Working Paper No. 132

October 2005

Martin Wittenberg is an Associate Professor in the School of Economics, University of Cape Town and a Research Associate at SALDRU.

Testing for a common latent variable in a linear regression: Or how to "fix" a bad variable by adding multiple proxies for it*

Martin Wittenberg
School of Economics
University of Cape Town

Abstract

We analyse models in which additional "controls" or proxies are included in a regression. This might occur intentionally if there is significant measurement error in a key regressor or if a key variable is not measured at all. We develop a test of the hypothesis that a subset of the regressors are all proxying for the same latent variable and we show how an estimate of the structural coefficient might be obtained more efficiently than is available in the current literature.

We apply the procedure to the determinants of sleep among young South Africans. We show that the income variable in the time use survey is badly measured. Nevertheless the measured impact of income on sleep is significant and amounts to 35 minutes per day between children with the median income and those in the topmost income bracket. Including a variety of asset proxies increases the estimated size of the coefficient enormously. The specification tests indicate that some of the asset proxies, however, have independent effects. Access to electricity, in particular, is not simply proxying for income. Instead it seems to be capturing access to various forms of entertainment, such as television. Even when this independent effect is properly accounted for, the size of the income coefficient is still 40% to 100% larger than in the specifications without the proxies.

Keywords: measurement error, proxy variables, specification tests

1 Introduction

Deliberately misspecifying a regression is hardly ever an optimal strategy. Textbooks deal with various types of nonintentional misspecification, the consequences flowing therefrom and ways of diagnosing the problem. In the most benign case, that of adding an irrelevant regressor, misspecification leads to a loss of efficiency. In other cases, such as the omission of a relevant variable or choice of incorrect functional form, the result of misspecification is inconsistency of the estimates. Measurement error is sometimes incorporated in discussions of specification and this problem too leads to inconsistent estimates. In this article we will suggest that there is an additional type of misspecification currently not recognised in the literature, that of including multiple versions of or proxies for one of the regressors. This

*I thank Murray Leibbrandt, Duncan Thomas, Chris Udry and participants at the 2004 African Econometric Society Conference for useful comments on this paper.

problem arises only if the regressor itself is mismeasured. We will introduce a test for this type of misspecification to determine whether it is plausible that several of the regressors are proxying for a common variable.

Generally such misspecification may be unintentional. However, in a recent article Lubotsky and Wittenberg (forthcoming) suggest that if multiple proxies are available for a regressor of interest, the attenuation bias resulting from the mismeasurement can be minimised by including all of the proxies in the main regression. This would amount to estimating an equation which, if interpreted structurally, would be misspecified, with the intention of estimating the structural parameters more accurately. The validity of the Lubotsky-Wittenberg (LW) estimation procedure hinges, of course, on the question whether it is plausible that these regressors are, indeed, all functioning as proxies for the same latent variable. Concerns about these issues may be well-founded, as we show in an example below. The test that we introduce can therefore also be seen as a test of the validity of the LW model. As such it should be of interest to researchers that have looked at this procedure (e.g. Bosworth and Collins 2003, Browning and Leth-Petersen 2003, Roland G. Fryer, Heaton, Levitt and Murphy 2005, Szalontai forthcoming). We show, furthermore, that identifying points at which the procedure breaks down can be very useful in revealing important "structural" relationships.

The attractiveness of the LW procedure is that in many empirical data sets key variables of interest are sometimes not measured at all. Indeed attempts to use multiple proxies or to combine them in indices have already straddled a number of areas of application (see, for instance, the literature that estimates "asset indices" on Demographic and Health Survey data as pioneered by Filmer and Pritchett (2001)). Even where the information is available, it may be of dubious quality. Income, for instance, is notoriously badly measured (for a discussion see Deaton 1997, pp.29–32). The impact of this misspecification on the estimates obtained from regressions including income is likely to be serious - both for the coefficient of income and, indeed, the coefficients of the more reliably measured variables. If the major

focus of interest is on the covariates, it may be tempting to augment the information obtained from the income variable by including various asset variables. We will show below that this strategy can, under certain circumstances, also yield "better" estimates of the coefficient on income.

In order to develop the intuition behind this result more precisely, consider the following regression model

$$y = x\beta + Z\gamma + \varepsilon \tag{1}$$

where x is a single variable and Z is a matrix of variables. In certain circumstances investigators are tempted to add additional variables. For instance, in many regressions "control variables" for location are routinely added to individual level regressions. Adding such regressors should not influence the estimation of the parameters in the model, if they are really superfluous. Suppose in particular that we wish to add x_1 where x_1 happens to be correlated with x, such that

$$x_1 = x\rho_1 + u_1$$

If x and Z are correctly measured and equation 1 is the structural model generating the data, then adding x_1 into the regression should lead to an insignificant coefficient on x_1. In particular if we estimate

$$y = xb_0 + x_1b_1 + Z\gamma + \varepsilon \tag{2}$$

then b_1 should be statistically zero. If it is not zero, then we would anticipate that x_1 belongs in the model.

The situation may, however, be more complicated if x is badly measured. If the measured variable is

$$x_0 = x + u_0 \tag{3}$$

then estimating the regression with x_0 in place of x may lead to **both** b_0 and b_1 being statistically significant, even when x_1 does not belong in the structural model. The proxy x_1

picks up part of the signal of the latent variable x. Lubotsky and Wittenberg (forthcoming) show that not only is it possible to estimate the model with both proxies for x included, but that the estimate of the coefficient on x, given by

$$\widehat{\beta} = \widehat{b}_0 + \widehat{\rho}_1 \widehat{b}_1$$

has lower attenuation bias than is achievable by any other linear combination of x_0 and x_1, including running the regression only on x_0.

This implies that in the presence of measurement error it would seem preferable to misspecify the regression by including additional proxies for the badly measured variable. Indeed this is precisely what we propose to do in the empirical section of this paper. We will show how adding in "asset variables" into a regression with a badly measured income variable can considerably strengthen the coefficient obtained on that income variable.

Nevertheless econometric "fixes" hardly ever come at zero cost. In this case problems may arise on two fronts. In the first instance, it may not be true that the asset variables are proxying for income. Instead the asset variables may very well belong in the regression themselves. As noted above, we cannot use tests of significance on the coefficients to decide this matter, because the asset variables may be significant, even though they are proxying for another variable.

Secondly, if any of the variables in the matrix Z are correlated with the "measurement error" terms u_0 and u_1, then the simple method of moments estimator for ρ_1 proposed by Lubotsky and Wittenberg (forthcoming) becomes inconsistent. Instead one should resort to the "covariates adjusted" estimator. Furthermore the estimated coefficient γ would be subject to an additional bias, over and above the "normal" bias induced by the measurement error.

Consider, for instance, the extreme case in which the structural model is given by

$$y = x\beta + z\gamma + \varepsilon$$

and the available variables are x_0, defined as in equation 3 and the "proxy" x_1 given by

$$x_1 = x\rho_1 + z$$

Estimating this regression with x_0, x_1 and z as explanatory variables, would yield expected coefficients of zero on x_0, $\frac{\beta}{\rho_1}$ on x_1 and $\gamma - \frac{\beta}{\rho_1}$ on z.

In short one should exercise due care when including additional proxies. The same point, of course, applies to the addition of "control variables". If these control variables are correlated with the variable of interest and there is measurement error in the system, then even a redundant variable can have a sizeable impact on the estimates.

This suggests that establishing whether a set of variables is proxying for a common latent variable would be useful in a number of different contexts. Most frequently it will be of interest to diagnose a new type of misspecification: interpreting the coefficient of a proxy variable as though it had a structural meaning, whereas it is simply reflecting part of the signal that is coming from a variable that is supposedly already included in the regression. Being able to diagnose this misspecification would aid in the proper interpretation of the coefficients.

In this paper we will introduce a set of procedures which can test for this type of misspecification under a broad set of conditions. In the process we will introduce a more efficient version of the Lubotsky-Wittenberg estimator. Indeed, the empirical examples below can most usefully be seen as extending that work.

The plan of the discussion is as follows. In the next section we introduce the central insight of this paper: that the estimator of ρ in the LW framework can best be thought of as a particular type of instrumental variable estimator. If there are covariates in the model it transpires that more instruments may become available. This in turn opens up the possibility of testing the validity of the moment conditions. We introduce this in section 3. First we show how the proxies can be tested individually and then, in sections 4 and 5 we do so for the system as a whole. This test of the overidentifying restrictions can be thought of

as an omnibus test for all sorts of failures of the model. Acceptance of the null hypothesis, i.e. of model validity, is equivalent to accepting that the variables tested are all proxying for a common latent variable.

In section 6 we demonstrate the procedures by means of an empirical example based on the South African Time Use Survey. The example shows some of the complexities in applying the technique as well as some of the potential.

2 The regression with proxy variables and one covariate

Let us consider the model

$$y = x\beta + z\gamma + \varepsilon$$

where we assume that z is measured accurately and $cov\,(x, z) = \sigma_{xz}$ is non-zero. We continue to assume that we have two proxy variables

$$\begin{aligned} x_0 &= x + u_0 \\ x_1 &= x\rho_1 + u_1 \end{aligned}$$

where we assume that the error variables u_j are uncorrelated with x, z and ε. The estimated model is

$$y = x_0 b_0 + x_1 b_1 + z\theta + \eta$$

We have selected a different symbol for the coefficient of z to indicate that typically this will not be (not even asymptotically) equal to γ.

As it stands there are no intercepts in this model. We have implicitly projected all variables on a constant, i.e. we are writing the model in deviations form. This means that all the errors will automatically have zero mean, so that $E\left(u_j u_j'\right) = \sigma_j^2 I_n$. Hence the movement (at various stages below) between writing the model in terms of covariances or the expected values of products.

The empirical information at our disposal is summarised in the following correlation matrix:

$$\begin{bmatrix} \beta^2\sigma_x^2 + 2\beta\gamma\sigma_{xz} + \gamma^2\sigma_z^2 + \sigma_\varepsilon^2 & \beta\sigma_x^2 + \gamma\sigma_{xz} & \beta\rho_1\sigma_x^2 + \gamma\rho_1\sigma_{xz} & \beta\sigma_{xz} + \gamma\sigma_z^2 \\ & \sigma_x^2 + \sigma_0^2 & \rho_1\sigma_x^2 + \sigma_{01} & \sigma_{xz} \\ & & \rho_1^2\sigma_x^2 + \sigma_1^2 & \rho_1\sigma_{xz} \\ & & & \sigma_z^2 \end{bmatrix}$$

We observe that ρ_1 is now **overdetermined**: $\rho_1 = cov(y, x_1)/cov(y, x_0)$ and $\rho_1 = cov(z, x_1)/cov(z, x_0)$ This raises the question as to how to estimate ρ_1 most efficiently.

Consideration of the form of the GMM estimates of ρ_1 given above, suggest that it can be thought of as an instrumental variables estimator in the regression of x_1 on x_0. We have

$$\begin{aligned} x_1 &= x\rho_1 + u_1 \\ &= (x_0 - u_0)\rho_1 + u_1 \\ &= x_0\rho_1 + \upsilon_1 \end{aligned}$$

Note that this cannot be estimated consistently by OLS, since x_0 is correlated with u_0 and hence υ_1. Since we have assumed that $cov(y, x_1) \neq 0$, and $cov(y, \upsilon_1) = 0$ (i.e. neither of the measurement error terms are correlated with any of the terms in the main regression), y is a legitimate instrument for x_1 in this regression. This may seem somewhat surprising given that y is actually the dependent variable in a regression in which x_1 is an explanatory variable. In addition note that if $cov(z, x_1) \neq 0$ and z is uncorrelated with any of the error terms, then z is also a legitimate instrument. This suggests that the optimal estimation strategy for ρ_1 is to use two-stage least squares, with y and z as instruments for x_0.

What happens if z is correlated with any of the u_j terms? In that case it would clearly be invalid to use z as an instrument. Note, however, that we can write

$$x_1 = x_0\rho_1 + z\phi_1 + \upsilon_1 \tag{4}$$

where by construction v_1 is now uncorrelated with z. We can estimate ρ_1 and ϕ_1 in the standard way, using y as an instrument for x_0 and z as an instrument for itself. The estimate of ρ_1 obtained in this way is numerically identical to the "covariate adjusted" estimator suggested in Lubotsky and Wittenberg (forthcoming), except that we do not have to obtain the residuals first.

One advantage of writing the proxy in the form of equation 4 is that it is possible to use the estimate of ϕ_1 to calculate a LW estimate of γ similar to the LW estimate of β. In this case the proxy variable adjusted estimate would be

$$\widehat{\gamma} = \widehat{\theta} + \widehat{\phi}_1 \widehat{b}_1$$

Although z is not mismeasured, the fact that b_1 is asymptotically nonzero due to the measurement error in x requires this LW adjustment to be made. With the adjustment, however, the overall bias in the estimate of γ would be lower than in any regression with some other linear function of x_0 and x_1, in particular a regression in which only x_0 is used as an explanatory variable. This follows, by extension, from the results in Lubotsky and Wittenberg (forthcoming).

3 Estimation and testing of ρ and ϕ proxy by proxy

We will now consider the more general model

$$\begin{aligned} y &= x\beta + Z\gamma + \varepsilon && \text{(5a)} \\ x_0 &= x + u_0 && \text{(5b)} \\ x_1 &= x\rho_1 + Z_{(1)}\phi_1 + u_1 && \text{(5c)} \\ &\cdots \\ x_k &= x\rho_k + Z_{(k)}\phi_k + u_k && \text{(5d)} \end{aligned}$$

where $cov(\varepsilon, u_j) = 0$. Furthermore we assume that $E(Z'u_0) = \mathbf{0}$. Since Z is assumed to contain a column of ones this implies that u_0 has mean zero. Consequently x_0 is an instance

of "classical measurement error". If this assumption is violated, the LW procedure can still be applied, but the coefficients of the Z variables that are correlated with u_0 will not be correctable by means of the procedure outlined below. We allow the other k proxies for x, viz. $x_1, \ldots, x_k$ to be more flexibly defined. In particular we suppose that the deviations from the latent variable x may be systematic and explicable in terms of some of the covariates.

We assume that the matrix of covariates Z can be partitioned as

$$Z = \begin{bmatrix} Z_{(i)} & Z_{-(i)} \end{bmatrix}$$

into variables $Z_{(i)}$ that have a direct effect on the proxy x_i, when controlling for the latent variable x, as well as variables $Z_{-(i)}$ that are correlated with x_i only through x. If there is a variable that is orthogonal to x and x_i we will include it in $Z_{(i)}$ to remove any ambiguity. It will generally be the case that $Z_{(i)}$ will contain at least a column of ones for the intercept. With constants in all equations we can assume that all error terms are mean zero. Finally we assume that $Z_{-(i)}$ is also non-empty. Note that in many applications including the ones reported on below we may be able to partition the matrix in the same way for every proxy.

With these assumptions, we can write the i-th proxy as

$$\begin{aligned} x_i &= x_0\rho_i + Z_{(i)}\phi_i + u_i - u_0\rho_i \\ &= M_{(i)}\delta_i + \upsilon_i \end{aligned} \tag{6}$$

where $M_{(i)} = \begin{bmatrix} x_0 & Z_{(i)} \end{bmatrix}$ and $\delta_i = \begin{bmatrix} \rho_i \\ \phi_i \end{bmatrix}$. Furthermore by our assumptions,

$$Z'\left(x_i - x_0\rho_i - Z_{(i)}\phi_i\right) = 0 \tag{7a}$$

$$y'\left(x_i - x_0\rho_i - Z_{(i)}\phi_i\right) = 0 \tag{7b}$$

and hence ρ_i and the coefficient vector ϕ_i can be consistently estimated by instrumental variables. In addition, provided that $Z_{-(i)}$ is non-empty our estimates are over-determined. An appropriate estimator for this equation would therefore be the generalised IV estimator,

or two-stage least squares estimator

$$\widehat{\delta}_i = \left(M'_{(i)} Z_y \left(Z'_y Z_y \right)^{-1} Z'_y M_{(i)} \right)^{-1} M'_{(i)} Z_y \left(Z'_y Z_y \right)^{-1} Z'_y x_i \tag{8}$$

where Z_y is the matrix of instruments, i.e.

$$Z_y = \left[\begin{array}{cc} y & Z \end{array} \right] \tag{9}$$

With these estimates for ρ_i and ϕ_i the LW estimates of β and γ will be given by

$$\widehat{\beta} = b_0 + \sum_{i=1}^{k} \widehat{\rho}_i b_i \tag{10a}$$

$$\widehat{\gamma}_j = \widehat{\theta}_j + \sum_{i=1}^{k} \widehat{\phi}_{ij} b_i \tag{10b}$$

where $\widehat{\theta}_j$ is the unadjusted coefficient on z_j in the multiple regression with all the proxies included and where $\widehat{\phi}_{ij} = 0$ if $z_j \notin Z_{(i)}$.

If $Z_{-(i)}$ is non-empty it is possible to test for the validity of the LW model by means of a test of the overidentifying restrictions. A particularly easy form of such a test is described in Davidson and MacKinnon (1993, p.236). They show that n times the uncentered R^2 from a regression of the IV residuals on the instruments Z is distributed as χ^2 with degrees of freedom equal to the degree of overidentification. In this case the degree of overidentification is exactly equal to the number of variables in $Z_{-(i)}$.

The null hypothesis for this test is that the moment conditions in equations 7 are all valid. This hypothesis could be rejected for a number of reasons:

1. ε is correlated with any of the u_i terms

2. one or more of the covariates in $Z_{-(i)}$ is correlated with any of the u_i terms

3. The proxy model is misspecified, i.e. it is not the case that

$$x_i = x_0 \rho_i + Z_{(i)} \phi_i + \upsilon_i$$

This could be due to either the fact that it is not the case that

$$x_i = x\rho_i + Z_{(i)}\phi_i + u_i$$

or that it is not the case that

$$x_0 = x + u_0$$

4. The main regression model is misspecified

Most of these would be reasons for being sceptical about the validity of applying the LW model. If the test fails for the second reason, however, it would be possible to re-partition the covariates and include the offending variables in the controls $Z_{(i)}$. If all covariates end up being included it is, of course, no longer possible to test the model. This may also be an indication that the LW model is of dubious value.

As an aside, we note that the "covariates adjusted" estimator of ρ suggested in Lubotsky and Wittenberg (forthcoming) is identical to the estimate obtained if $Z_{(i)} = Z$, i.e. if none of the covariates was a legitimate instrument. We would expect therefore that the procedure outlined above should lead to more efficient estimates than that given in the original paper.

4 Systems estimation of ρ and ϕ

According to the model given in equations 5 there are k equations of the type 6 to estimate. We can "stack" these equations in the standard way:

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} = \begin{bmatrix} M_{(1)} & 0 & \cdots & 0 \\ 0 & M_{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & M_{(k)} \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_k \end{bmatrix} + \begin{bmatrix} \upsilon_1 \\ \upsilon_2 \\ \vdots \\ \upsilon_k \end{bmatrix} \tag{11}$$

$$\mathbf{x}_v = \mathbf{M}\boldsymbol{\delta} + \boldsymbol{\upsilon}_v \tag{12}$$

In the special case where the same control variables $Z_{(i)} = Z_u$ are used in every equation the matrix $\mathbf{M}$ takes on the particularly simple form

$$\begin{aligned} \mathbf{M} &= I_k \otimes \begin{bmatrix} x_0 & Z_u \end{bmatrix} \\ &= I_k \otimes M_x \end{aligned} \tag{13}$$

where $M_x = \begin{bmatrix} x_0 & Z_u \end{bmatrix}$.

The systems version of the proxy by proxy estimation outlined above would be to define the matrix of instruments

$$\begin{aligned} \mathbf{Z}_v &= \begin{bmatrix} Z_y & 0 & \cdots & 0 \\ 0 & Z_y & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Z_y \end{bmatrix} \\ &= I_k \otimes Z_y \end{aligned} \tag{14}$$

and then define the "2SLS" estimator

$$\widehat{\boldsymbol{\delta}}_{2SLS} = \left(\mathbf{M}'\mathbf{Z}_v \left(\mathbf{Z}_v'\mathbf{Z}_v\right)^{-1} \mathbf{Z}_v'\mathbf{M}\right)^{-1} \mathbf{M}'\mathbf{Z}_v \left(\mathbf{Z}_v'\mathbf{Z}_v\right)^{-1} \mathbf{Z}_v'\mathbf{x}_v \tag{15}$$

which is numerically equal to the proxy by proxy estimation of $\boldsymbol{\delta}$ given in equation 8.

This estimator, however, may not be efficient, since the covariance matrix of $\boldsymbol{\upsilon}_v$ is not diagonal. It is obvious that the vector υ_i and the vector υ_j are correlated. In fact a typical covariance between an element of υ_i and υ_j will be given by $cov\left(\upsilon_i, \upsilon_j\right) = cov\left(u_i - u_0\rho_i, u_j - u_0\rho_j\right) = \sigma_{ij} - \rho_j\sigma_{0i} - \rho_i\sigma_{0j} + \rho_i\rho_j\sigma_0^2$. Let us denote this as υ_{ij}. Then

$$\begin{aligned} E\left(\boldsymbol{\upsilon}_v\boldsymbol{\upsilon}_v'\right) &= \Psi \\ &= \Sigma \otimes I_n \end{aligned}$$

where

$$\Sigma = \begin{bmatrix} \upsilon_{11} & \upsilon_{12} & \cdots & \upsilon_{1k} \\ \upsilon_{12} & \upsilon_{22} & \cdots & \upsilon_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \upsilon_{1k} & \upsilon_{2k} & \cdots & \upsilon_{kk} \end{bmatrix}$$

This suggests that the system of proxy equations could be more efficiently estimated as a system, taking these cross-equation correlations into account.

A brief consideration of the formal properties of the model shows that the GLS estimator of the system is identical to the three stage least squares estimator described in the literature (Davidson and MacKinnon 1993, Mittelhammer, Judge and Miller 2000). This means that estimation of $\boldsymbol{\delta}$ and hence ρ can proceed with readily available software. Furthermore the asymptotic properties of this estimator are well established.

Nevertheless it is useful to briefly go through the derivation of the estimator, since we will use some of the intermediate results in section 5. We follow the treatment in Mittelhammer et al. (2000, pp.462–465) and develop the estimator as the optimal GMM estimator from the population moment condition

$$E\left(\mathbf{Z}_v'\left(\mathbf{x}_v - \mathbf{M}\boldsymbol{\delta}\right)\right) = \mathbf{0} \tag{16}$$

This yields the sample counterpart

$$\frac{1}{n}\mathbf{Z}_v'\left(\mathbf{x}_v - \mathbf{M}\boldsymbol{\delta}\right) = \mathbf{0}$$

and hence the GMM estimator

$$\widehat{\boldsymbol{\delta}}\left(\mathbf{W}\right) = \left[\mathbf{M}'\mathbf{Z}_v\mathbf{W}\mathbf{Z}_v'\mathbf{M}\right]^{-1}\mathbf{M}'\mathbf{Z}_v\mathbf{W}\mathbf{Z}_v'\mathbf{x}_v \tag{17}$$

for some positive definite weighting matrix $\mathbf{W}$. Indeed with $\mathbf{W} = \left(\mathbf{Z}_v'\mathbf{Z}_v\right)^{-1}$ we get the

"2SLS" estimator of equation 15. The **optimal** weighting matrix, however, is

$$\begin{aligned} w_*^{-1} &= cov\left(n^{-\frac{1}{2}}\mathbf{Z}_v'\boldsymbol{v}_v\right) \\ &= n^{-1}E\left(\mathbf{Z}_v'\boldsymbol{v}_v\boldsymbol{v}_v'\mathbf{Z}_v\right) \\ &= n^{-1}E\left(\left(I_k\otimes Z_y\right)'\left(\Sigma\otimes I_n\right)\left(I_k\otimes Z_y\right)\right) \\ &= \Sigma\otimes E\left(n^{-1}Z_y'Z_y\right) \end{aligned} \tag{18}$$

$\mathbf{W}$ is therefore proportional to $\Sigma^{-1}\otimes\left(Z_y'Z_y\right)^{-1}$ and the corresponding GMM estimator is given by

$$\boldsymbol{\delta}_{GMM} = \left[\mathbf{M}'\left(\Sigma^{-1}\otimes Z_y\left(Z_y'Z_y\right)^{-1}Z_y'\right)\mathbf{M}\right]^{-1}\mathbf{M}'\left(\Sigma^{-1}\otimes Z_y\left(Z_y'Z_y\right)^{-1}Z_y'\right)\mathbf{x}_v \tag{19}$$

The matrix Σ is unknown, but can be consistently estimated using the residuals from any consistent GMM estimator. The 2SLS estimator (15) is particularly convenient in this regard. We have

$$\widehat{\Sigma}_{ij} = \frac{1}{n}\widehat{v}_i'\widehat{v}_j$$

where $\widehat{v}_i$ and $\widehat{v}_j$ are the vectors of residuals from the i-th and j-th proxy estimation respectively. Using this in place of Σ gives the estimated weighting matrix

$$\widehat{\mathbf{W}} = \widehat{\Sigma}^{-1}\otimes\left(Z_y'Z_y\right)^{-1} \tag{20}$$

and correspondingly the estimated optimal GMM estimator $\widehat{\boldsymbol{\delta}}_{GMM}$ or three-stage least squares estimator.

In the case where the same control variables $Z_{(i)} = Z_u$ are used in every equation, we can substitute equation 13 into equation 19. This yields

$$\begin{aligned} \widehat{\boldsymbol{\delta}}_{GMM} &= \left[\widehat{\Sigma}^{-1}\otimes M_x'Z_y\left(Z_y'Z_y\right)^{-1}Z_y'M_x\right]^{-1}\left[\widehat{\Sigma}^{-1}\otimes M_x'Z_y\left(Z_y'Z_y\right)^{-1}Z_y'\right]\mathbf{x}_v \\ &= \left\{I_k\otimes\left(M_x'Z_y\left(Z_y'Z_y\right)^{-1}Z_y'M_x\right)^{-1}M_x'Z_y\left(Z_y'Z_y\right)^{-1}Z_y'\right\}\mathbf{x}_v \end{aligned} \tag{21}$$

This, however, is numerically identical to the proxy by proxy estimation of $\boldsymbol{\delta}$ by 2SLS! This finding parallels the result that least squares estimation and GLS estimation of a SUR system with **identical** X matrices are equal (see for instance Mittelhammer et al. 2000, p.453).

5 Specification testing

Just as it is possible to test the overidentifying restrictions proxy by proxy, it is possible to do so on the system as a whole. The appropriate test statistic in this case will be given by

$$\left[\mathbf{Z}_v'\left(\mathbf{x}_v - \mathbf{M}\widehat{\boldsymbol{\delta}}_{GMM}\right)\right]'\widehat{\mathbf{W}}\left[\mathbf{Z}_v'\left(\mathbf{x}_v - \mathbf{M}\widehat{\boldsymbol{\delta}}_{GMM}\right)\right] \xrightarrow{d} \chi^2(d)$$

(Mittelhammer et al. 2000, pp.438–9) where $\widehat{\mathbf{W}}$ is given by equation 20 and d is the degree of overidentification, i.e. the number of instruments used in the system as a whole minus the number of moment equations. The test statistic τ can be simplified to

$$\widehat{v}_v'\left(\widehat{\Sigma}^{-1}\otimes Z_y\left(Z_y'Z_y\right)^{-1}Z_y'\right)\widehat{v}_v$$

where $\widehat{v}_v$ is the vector of stacked IV residuals. Now note that this is equivalent to

$$\widehat{v}_v'\left(I_k\otimes Z_y\left(Z_y'Z_y\right)^{-1}Z_y'\right)'\left(\widehat{\Sigma}^{-1}\otimes I_n\right)\left(I_k\otimes Z_y\left(Z_y'Z_y\right)^{-1}Z_y'\right)\widehat{v}_v$$

Let

$$\widehat{\widehat{v}} = \left(I_k\otimes Z_y\left(Z_y'Z_y\right)^{-1}Z_y'\right)\widehat{v}_v$$

This is the vector of fitted values from the artificial regression of the residuals $\widehat{v}_v$ on the matrix of instruments $\mathbf{Z}_v$. Hence the test statistic can also be written as

$$\widehat{\widehat{v}}'\left(\widehat{\Sigma}^{-1}\otimes I_n\right)\widehat{\widehat{v}}$$

In the trivial case where this estimator is applied to a "system" of only one equation $\widehat{\Sigma}^{-1} = \widehat{\sigma}^{-2}$ and the test statistic is identical to nR_u^2 in the regression of $\widehat{v}_v$ on $\mathbf{Z}_v$. Whenever $\widehat{\Sigma}^{-1}\otimes I_n$ is **not** equal to $\widehat{\sigma}^{-2}I_{kn}$, however, this statistic will not be computable in this manner. Instead,

let

$$\widehat{\Sigma}^{-1} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1k} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{k1} & \alpha_{k2} & \cdots & \alpha_{kk} \end{bmatrix}$$

and let $\widehat{\widehat{v}}_i$ refer (in the obvious way) to the vector of fitted values corresponding to the i-th proxy then

$$\begin{aligned} \widehat{\widehat{v}}' \left(\widehat{\Sigma}^{-1} \otimes I_n\right) \widehat{\widehat{v}} &= \begin{bmatrix} \widehat{\widehat{v}}_1' & \widehat{\widehat{v}}_2' & \cdots & \widehat{\widehat{v}}_k' \end{bmatrix} \begin{bmatrix} \alpha_{11}I_n & \alpha_{12}I_n & \cdots & \alpha_{1k}I_n \\ \alpha_{21}I_n & \alpha_{22}I_n & \cdots & \alpha_{2k}I_n \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{k1}I_n & \alpha_{k2}I_n & \cdots & \alpha_{kk}I_n \end{bmatrix} \begin{bmatrix} \widehat{\widehat{v}}_1 \\ \widehat{\widehat{v}}_2 \\ \vdots \\ \widehat{\widehat{v}}_k \end{bmatrix} \\ &= \begin{bmatrix} \sum_{i=1}^k \alpha_{i1}\widehat{\widehat{v}}_i' & \sum_{i=1}^k \alpha_{i2}\widehat{\widehat{v}}_i' & \cdots & \sum_{i=1}^k \alpha_{ik}\widehat{\widehat{v}}_i' \end{bmatrix} \begin{bmatrix} \widehat{\widehat{v}}_1 \\ \widehat{\widehat{v}}_2 \\ \vdots \\ \widehat{\widehat{v}}_k \end{bmatrix} \\ &= \sum_{i=1}^k \sum_{j=1}^k \alpha_{ij} \widehat{\widehat{v}}_i' \widehat{\widehat{v}}_j \end{aligned} \tag{22}$$

If we define the matrix B as

$$B_{ij} = \widehat{\widehat{v}}_i' \widehat{\widehat{v}}_j$$

and we note that B is symmetric, it is easy to see that

$$\sum_{i=1}^k \sum_{j=1}^k \alpha_{ij} \widehat{\widehat{v}}_i' \widehat{\widehat{v}}_j = tr\left(\widehat{\Sigma}^{-1} B\right)$$

If the artificial regression fits perfectly, so that $\widehat{\widehat{v}}_j = \widehat{v}_j$, then $B = n\widehat{\Sigma}$ and the test statistic will be kn.

The important point to note is that the calculation of the test statistic does not require the formation of any $kn \times kn$ matrices. Indeed provided that one uses the same partition

of the Z matrix in each proxy equation there is no need to compute anything other than the proxy-by-proxy two stage least squares estimates and the residuals from them. Even in this case one may want to utilise three stage least squares routines since these conveniently calculate the matrix $\widehat{\Sigma}$. Certainly one will need to do so if one wanted to test restrictions on the $\boldsymbol{\rho}$ vector.

In the case where the same controls are used in each proxy equation it is easy to calculate the degrees of freedom for the hypothesis test. If we let the number of variables that are **not** in Z_u be l, i.e.

$$l = rank\left(Z\right) - rank\left(Z_u\right)$$

then the degrees of freedom for the test will be $d = lk$, i.e. the number of instruments $\left(rank\left(Z\right)+1\right)k$ used less the number of moment equations, i.e. $\left(1+rank\left(Z_u\right)\right)k$.

The null hypothesis for this test is that the moment condition given in equation 16 holds. As we noted above, a rejection of the null can occur for a number of different reasons:

- A correlation between ε and any of the u_i terms. In this case y would cease to be an appropriate instrument. This implies that the proxy variable x_i is not proxying only for the latent variable x, but should be in the main regression.
- A correlation between any of the z_j variables used as instruments and any of the u_i terms. In this case z_j should be treated as a control in equation i and not as an instrument.
- A misspecification of any of the proxy variable equations
- A misspecification of the main regression

In short rejection of the null hypothesis should probably be taken as evidence that the model should be seriously rethought, although it may be technically possible to respecify it in ways that provide consistent estimates of ρ.

6 An application: the demand for sleep among schoolgoing South Africans

In order to explore how these techniques work on real data, we apply them to the problem of estimating the impact of household income on time spent sleeping by young South Africans. The path-breaking study on the economic determinants of sleep was by Biddle and Hamermesh (1990). They argued that the length of sleep was not purely biologically determined but responded to economic signals. The opportunity cost of sleep is the wage foregone and so one would expect that high wage earners sleep less than low wage ones. They show this on US time use data. Szalontai (forthcoming) finds a similar relationship using South African data. A recent study of time use among South African school age children found that sleep decreases with total household income (Wittenberg 2005). In this case the opportunity cost of sleep cannot be the wage, so there must be other explanations, including the entertainment opportunities foregone. Perhaps sleep is just an inferior good!

The data for the South African studies comes from the South African time use survey conducted by Statistics South Africa (Budlender, Chobokoane and Mpetsheni 2001), a cross-sectional survey of approximately 14 000 individuals in about 8 000 households. The time use information is contained in a 24-hour recall diary, organised into 30 minute intervals. During each interval individuals could indicate up to three activities that they engaged in. Activities were coded in terms of a standard activity classification system.

One of the problems with investigating economic relationships on this data set is that the income information is very poor. In this it is not unique: in all surveys there is some trade off between the breadth of issues covered and the quality of information obtained on particular variables. Surveys which ask a simple income question and do not extensively probe are unlikely to get quality income data. This has been shown in the case of South Africa's census data by Alderman et al. (2000).

The extent of the problem can be seen from Table 1, which contrasts the total household income distribution from the Time Use Survey and the expenditure and income distributions

from an Income and Expenditure Survey conducted by Statistics South Africa in the same year, i.e. 2000.

It is clear that total household income in the Time Use Survey has been measured with considerable error. It also seems clear that this error is not classical measurement error in the form given in equation 5b. There seems to be a systematic underreporting of incomes, so that the measurement error term u_0 has a non-zero mean. This means that the intercept term in the regression will have an additional unknown bias. Furthermore it is plausible that the scale of the measurement error may be correlated with some of the covariates. The LW "corrected" coefficients of these terms should therefore be viewed with some caution.

The scale of the underreporting however also increases the attractiveness of the LW procedure. It seems clear that households with very low reported incomes, but possessing fridges, stoves and televisions must be better off than households without those. Including asset proxies would control for income much better than using the reported income categories alone.

For the purposes of the regressions we have turned the discrete income categories into a continuous variable, by taking the midpoints of the categories and twice the lower bound of the open category. The latter is recommended by Charles Simkins (personal communication), based on the fact that the income distribution at the top end is roughly Pareto with parameter just under two. This adds an additional level of noise to the variable which will, however, be of secondary influence compared to the underreporting shown above.

In column 1 of Table 2 we report a simple regression of minutes spent sleeping during an average night during the school week (Monday to Friday). The sample is restricted only to individuals who are actually observed attending school in the time diaries. The coefficient on log income is around -10, which would translate to a 35 minute difference between children with the median income (R600) and the richest children (R20 000). Adding in the proxies we see a marked drop in this coefficient, but significant coefficients on "fridge" and "tv". The LW estimates in columns (3) through (5) make different assumptions about which covariates

are correlated with the proxies and which are independent of the measurement errors. The estimates in column (3) correspond to the "covariates adjusted" LW estimator suggested in Lubotsky and Wittenberg (forthcoming). This estimator gives by far the largest point estimate, almost four times the size of the original coefficient in column (1). It is impossible to subject this estimator to a specification test, since this model is exactly identified. We note, however, that in line with the poor finite sample performance of just-identified instrumental variable estimates (Davidson and MacKinnon 1993), the standard errors are much larger than is the case with the other estimates.

We can subject the other two models to specification tests. In column (4) we report not only the LW coefficients (in the column labelled "b") but also the estimates of ρ, the overidentification test statistics τ for each of the proxy equations as well as the associated degrees of freedom and p-values. In this case it is clear that the model is overwhelmingly rejected by the data. The covariates are not independent of the proxies, or some of the proxies belong in the main regression.

In column (5) we have included household size, the number of children in the household and a set of racial dummies as covariates that are likely to be correlated with the proxies. The remaining covariates are attributes of the child as well as the day of the week and month of the year during which the survey was conducted. It is less plausible that these should be correlated with the assets. Indeed most of the proxy equations would now accept the null hypothesis that the asset in question is proxying for income. However we would still reject the hypothesis that ownership of a television is proxying for income. The systems test would, however, accept that the LW model as a whole is reasonable, at the 5% level.

Nevertheless the coefficient of -27 is perhaps still too large to be plausible. It would correspond to a one-and-a-half hour difference in sleep time between the children with median reported incomes and the richest children. Furthermore the rejection of the overidentification test in the television equation gives pause for thought.

In column (6) we report a simple OLS regression with income and television ownership

considered as separate regressors. The coefficient on income shrinks (although it is still statistically non-zero) while television makes a noticeable difference to sleeping times. The LW estimates corresponding to this regression are given in column (7). They give a doubling of the income coefficient compared to column (6) with a small reduction in the size of the television effect. The point estimate on income corresponds to about 48 minutes difference in sleep times between the median income and the top. This seems a more plausible impact, particularly if one bears in mind the additional 23 minutes loss of sleep attendant on owning a television!

As this regression performs well on all the specification tests we might be happy to leave the issue. There seem to be two factors that matter: income raises the opportunity costs of sleep, while ownership of a television does so even when controlling for income.

This simple picture is, however, muddied by the results shown in Table 3. In this table we have added in two "infrastructure" proxies - use of electricity for lighting and living in a brick dwelling. The LW estimates reported in column (5) correspond to the same model as that accepted by the data in column (7) of Table 1, except that these additional proxies have been used. In this case, however, the specification test roundly rejects the validity of the model. More particularly it suggests that electricity should be in the main model. It also raises the prospect that perhaps it is not television *per se* that is important for sleep, but access to electricity and the attendant opportunities. Indeed in looking at the results in column (5) of Table 1 we note that the data suggest that both access to a fridge and access to a stove may be important for sleep over and above income.

In Table 3 we explore two competing hypotheses:

- In columns (1) and (5) we have the hypothesis that only the log of income and television ownership belong in the structural equation. Access to electricity is interpreted in the LW model as just another proxy for mismeasured income. As we have noted above, this model is not supported by the specification tests - either for the "electricity" proxy equation or for the system as a whole.

- In columns (2), (6) and (7) we explore the hypothesis that only the log of income and access to electricity belong in the structural equation. TV ownership is relegated to the status of a proxy for income. This model is accepted by the data at conventional levels of significance. The difference between the LW estimates in columns (6) and (7) is that in the latter case we have used only TV ownership as proxy, i.e. the LW corrections were applied to the "raw" regression shown in column (3). Since there was only one proxy equation, the systems version of the overidentification test is numerically equal to the single equation test. The increase in the coefficient of log income between columns (7) and (6) suggest that the additional proxies do add some information, although they do so with considerable noise. The bootstrap standard errors are much larger with all of the proxies included, suggesting that perhaps one might prefer the estimates contained in column (7).

The procedure that we have employed in adjudicating whether television or electricity are structural is reminiscent of Granger causality tests. Like those tests it is possible for the answers to be less clear cut. The data might have suggested that both variables belong in the main regression or that neither do. Of course tests are hardly ever completely decisive and the power of the tests may be inadequate. For the sceptical reader we provide the results for the model in which both television and electricity are interpreted as belonging in the structural equation. Comparing the estimates in column (3) with the corresponding LW estimates in column (8) we note that the simple OLS results probably understate the importance of income. Some of the effects of income are captured by the coefficients of television and, to a lesser extent, electricity.

Reviewing the evidence available in Tables 2 and 3 we come to the conclusion that a point estimate of around -15 (Table 3, column 7) probably represents the best guess at the impact of income on sleep. This would correspond to a difference of 52 minutes of sleep per day between kids at the median income versus children in the top income bracket. In addition to this children with electrified houses would sleep 24 minutes a day less. One of the

opportunities opened up by larger incomes is, of course, the ability to watch television. So television probably matters - but as an outcome rather than a structural determinant. Indeed an initial look at patterns of television watching reported in Wittenberg (2005) suggests that television viewing may be related to income in inverse U fashion - increasing with income and then decreasing at the top. The richest children probably have many more non-TV opportunities (e.g. internet chat rooms) to while away the nights. In short there are non-technical reasons for believing that the specification tests have picked out perhaps the most plausible model of all.

7 Conclusion

The Lubotsky-Wittenberg procedure (forthcoming) is designed to provide a framework for thinking about the relationships between proxies and the underlying latent variables. In particular it allows one to estimate the structural coefficients in a regression analysis. In this paper we have shown how one might estimate the model more efficiently and test for its validity. The empirical application suggests that these tests may be rather good at rejecting inappropriate specifications and picking out more plausible ones. Along the way we have shown that the tests can be used also to adjudicate which of two proxies is "more structural" than the other. Ultimately these techniques are no substitute for thinking about the underlying relationships theoretically. We might have accepted the simple model with only log of income in the main regression on the basis of the omnibus systems test. The point estimate of -27 is pleasingly large, except perhaps implausibly so. By contrast a different author might be convinced on the basis of the fact that television ownership is statistically significant in every one of the proxy regressions that it really is structural.

In short, the LW procedure cannot substitute for proper judgement. It also cannot in any real way "fix" bad data. There is no substitute for quality information. Nevertheless it can be highly suggestive. Depending on which model one prefers, the addition of the proxies leads from a 40% to a 100% (or more!) increase in the size of the coefficients. This sort of

difference is sizeable in the context of the models analysed.

References

Alderman, Harold et al., "Combining census and survey data to construct a poverty map of South Africa," in "Measuring poverty in South Africa," Statistics South Africa, 2000.

Biddle, Jeff E. and Daniel S. Hamermesh, "Sleep and the Allocation of Time," *Journal of Political Economy*, 1990, *98* (5 Part 1), 922–943.

Bosworth, Barry P. and Susan M. Collins, *Brookings Papers on Economic Activity*, 2003, *2003* (2), 113–179.

Browning, Martin and Soren Leth-Petersen, "Imputing Consumption from Income and Wealth Information," *Economic Journal*, 2003, *113*, F282–F301.

Budlender, Debbie, Ntebaleng Chobokoane, and Yandiswa Mpetsheni, *A Survey of Time Use: How South African women and men spend their time*, Pretoria: Statistics South Africa, 2001.

Davidson, Russell and James G. MacKinnon, *Estimation and Inference in Econometrics*, New York: Oxford University Press, 1993.

Deaton, Angus, *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*, Baltimore: Johns Hopkins University Press, 1997.

Filmer, Deon and Lant H. Pritchett, "Estimating Wealth Effects Without Expenditure Data – Or Tears: An Application to Educational Enrollment in States of India," *Demography*, February 2001, *38* (1), 115–132.

Fryer, Jr. Roland G., Paul S. Heaton, Steven D. Levitt, and Kevin M. Murphy, "Measuring the Impact of Crack Cocaine," Working Paper 11318, NBER May 2005.

Lubotsky, Darren and Martin Wittenberg, "Interpretation of regressions with multiple proxies," *Review of Economics and Statistics*, forthcoming.

Mittelhammer, Ron C., George G. Judge, and Douglas J. Miller, *Econometric Foundations*, Cambridge: CUP, 2000.

Szalontai, Gabor, "The Demand for Sleep: A South African Study," *Economic Modelling*, forthcoming.

Wittenberg, Martin, "The School Day in South Africa," Working Paper 113, CSSR, University of Cape Town 2005.

Table 1: Comparing Income distributions in the IES and the Time Use Survey

Category (per month)	Income and Expenditure Survey 2000 Expenditure	Income and Expenditure Survey 2000 Income (proportions)	Time Use Survey 2000 Income
R 0-399	0.076	0.097	0.204
R 400-799	0.195	0.215	0.286
R 800-1199	0.160	0.144	0.141
R 1200-1799	0.151	0.134	0.103
R 1800-2499	0.105	0.092	0.068
R 2500-4999	0.151	0.149	0.093
R 5000-9999	0.089	0.090	0.074
R 10000+	0.073	0.079	0.031
	1.000	1.000	1.000

Table 2: The determinants of sleep among schoolgoing adolescents in South Africa

Sleep	Without proxies **(1)**	With asset proxies **(2)**	LW - All covariates used as controls **(3)**		LW - All covariates deemed exogenous **(4)**					LW - Some covariates deemed exogenous **(5)**					Income and TV **(6)**	LW - using asset proxies, TV ownership included as covariate **(7)**				
N=1867	b	b	b	ρ	b	ρ	τ	df	p value	b	ρ	τ	df	p value	b	b	ρ	τ	df	p value
loginc	-9.98***	-4.05+	-36.11		-18.49					-27.05					-6.8**	-13.64997				
	(2.14)	(2.36)	(11.04)		(4.12)					(5.09)					(2.18)	(4.16)				
washmach		-1.35		0.132		0.357	91.65	25	0.000		0.176	8.79	9	0.457			0.183	9.45	9	0.397
		(7.59)		(0.07)		(0.014)					(0.037)						(0.047)			
vacuum		-4.16		0.123		0.296	129.92	25	0.000		0.172	8.81	9	0.455			0.198	8.06	9	0.528
		(8.47)		(0.063)		(0.016)					(0.042)						(0.054)			
fridge		-10.98*		0.748		0.335	138.84	25	0.000		0.567	15.08	9	0.089			0.4	14.24	9	0.114
		(4.95)		(0.186)		(0.017)					(0.083)						(0.091)			
phone		-3		0.417		0.329	116.70	25	0.000		0.379	9.88	9	0.361			0.306	9.44	9	0.398
		(4.54)		(0.132)		(0.016)					(0.065)						(0.08)			
stove		-7.38		0.625		0.326	166.67	25	0.000		0.423	15.96	9	0.068			0.268	14.39	9	0.109
		(4.73)		(0.165)		(0.017)					(0.076)						(0.086)			
tv		-18.02***		0.828		0.271	144.50	25	0.000		0.547	17.14	9	0.047	-26.32***	-23.48				
		(4.59)		(0.203)		(0.017)					(0.087)				(4.09)	(4.63)				
ownradio		-5.9		0.328		0.087	77.87	25	0.000		0.215	14.55	9	0.104			0.143	14.03	9	0.121
		(4.66)		(0.131)		(0.014)					(0.065)						(0.082)			
car		-0.45		0.262		0.285	62.02	25	0.000		0.179	9.73	9	0.372			0.147	10.65	9	0.300
		(4.93)		(0.11)		(0.016)					(0.06)						(0.076)			
clock		-1.18		0.267		0.127	80.93	25	0.000		0.311	5.61	9	0.778			0.281	5.70	9	0.769
		(4.56)		(0.114)		(0.015)					(0.071)						(0.089)			
hhsize	-0.06	0.91	2.50		see (2)					1.51					0.54	1.06				
	(1.34)	(1.33)	(1.95)							(1.40)					(1.32)	(1.56)				
hhkids	-0.17	-1.36	-2.78		see (2)					-1.70					-0.82	-1.29				
	(1.89)	(1.88)	(2.53)							(1.98)					(1.87)	(2.07)				
Coloured	-2.74	1.91	10.62		see (2)					6.35					-1.41	1.86				
	(8.41)	(8.66)	(12.12)							(9.68)					(8.32)	(10.07)				
Indian	13.18	16.31	42.45		see (2)					32.66					13.85	21.11				
	(12.89)	(13.04)	(16.11)							(13.61)					(12.75)	(12.53)				
White	13.71	19.02+	65.78		see (2)					48.21					13.31	26.42				
	(9.75)	(11.4)	(22.73)							(13.86)					(9.64)	(13.20)				
Controls for stratum and province	Y	Y	Y		see (2)					Y					Y	Y				
age	-3.9***	-4.83***	-5.71		see (2)					see (2)					-4.5***	see (6)				
	(1.1)	(1.1)													(1.1)					
educ	-5.13***	-3.98**	-3.17		see (2)					see (2)					-4.49***	see (6)				
	(1.21)	(1.21)													(1.2)					
gender	-4.74	-3.42	-4.95		see (2)					see (2)					-3.85	see (6)				
	(3.6)	(3.56)													(3.56)					
Controls for tranche and day of week	Y	Y	Y		see (2)					see (2)					Y	see (6)				
System Over-identification test							909.83	225	0.000			98.57	81	0.090				83.638	72	0.1643

Notes Standard errors are given in parentheses and are corrected for clustering. Standard errors for the LW estimates were calculated by means of a clustered bootstrap with 200 replications.
Significance level: + 10% * 5% ** 1% *** 0.1% (indicated only for OLS results)

Table 3: The determinants of sleep among schoolgoing adolescents in South Africa - the impact of TV and electricity

Sleep	TV	Electricity	TV and Electricity	Asset & infra-structure proxies	LW - TV as a covariate					LW - Electric lighting as a covariate						LW - TV & electricity as covariates				
	(1)	**(2)**	**(3)**	**(4)**	**(5)**					**(6)**					**(7)**	**(8)**				
N=1867	b	b	b	b	b	*ρ*	*τ*	df	p value	b	*ρ*	*τ*	df	p value	b	b	*ρ*	*τ*	df	p value
loginc	-6.8**	-8.25***	-6.22**	-4.22+	-13.43					-19.62					-15.64	-10.75				
	(2.18)	(2.14)	(2.17)	(2.36)	(4.41)					(4.58)					(3.67)					
washmach				-2.04		0.183	9.45	9	0.397		0.178	8.98	9	0.439			0.182	9.54	9	0.389
				(7.58)		(0.047)					(0.04)						(0.048)			
vacuum				-4.57		0.198	8.06	9	0.528		0.182	8.31	9	0.504			0.201	7.90	9	0.544
				(8.45)		(0.054)					(0.045)						(0.055)			
fridge				-7.78		0.4	14.24	9	0.114		0.46	14.81	9	0.096			0.359	14.81	9	0.096
				(5.08)		(0.091)					(0.078)						(0.086)			
phone				-2.04		0.306	9.44	9	0.398		0.333	9.24	9	0.415			0.288	9.10	9	0.428
				(4.57)		(0.08)					(0.068)						(0.08)			
stove				-2.92		0.268	14.39	9	0.109		0.285	8.29	9	0.506			0.209	7.62	9	0.573
				(4.95)		(0.086)					(0.068)						(0.078)			
tv	-26.32***		-20.5***	-16.3***	-23.52						0.459	14.62	9	0.102		-18.81				
	(4.09)		(4.29)	(4.63)	(4.56)						(0.087)									
ownradio				-6.22		0.143	14.03	9	0.121		0.211	14.59	9	0.103			0.146	13.94	9	0.125
				(4.66)		(0.082)					(0.07)						(0.083)			
car				-1.15		0.147	10.65	9	0.300		0.17	9.70	9	0.375			0.146	10.64	9	0.301
				(4.94)		(0.076)					(0.066)						(0.077)			
clock				-1.06		0.281	5.70	9	0.769		0.295	5.52	9	0.787			0.275	5.63	9	0.776
				(4.56)		(0.089)					(0.076)						(0.09)			
electric lights		-26.13***	-19.38***	-15.5**		0.143	29.12	9	0.001	-23.17					-24.22	-18.79				
		(4.29)	(4.49)	(5.06)		(0.092)				(4.67)					(4.89)					
brick dwelling				0.96		0.164	11.08	9	0.270		0.2	12.07	9	0.209			0.135	11.33	9	0.254
				(4.13)		(0.096)					(0.082)						(0.096)			
hhsize	0.54	0.41	0.76	1.01	1.04					1.39					1.06	1.10				
	(1.32)	(1.32)	(1.32)	(1.33)	(1.46)					(1.48)					(1.35)					
hhkids	-0.82	-0.98	-1.27	-1.61	-1.29					-1.89					-1.61	-1.57				
	(1.87)	(1.87)	(1.86)	(1.88)	(2.02)					(2.09)					(1.95)					
Coloured	-1.41	-1.76	-0.98	1.27	1.87					4.05					2.03	1.16				
	(8.32)	(8.33)	(8.29)	(8.65)	(10.88)					(10.51)					(9.91)					
Indian	13.85	17.48	16.89	18.6	21.35					29.45					25.30	21.46				
	(12.75)	(12.78)	(12.71)	(13.05)	(11.30)					(12.51)					(11.91)					
White	13.31	14.94	14.31	19.74+	26.22					37.23					29.45	22.85				
	(9.64)	(9.66)	(9.6)	(11.38)	(12.89)					(13.16)										
Controls for stratum and province	Y	Y	Y	Y	Y					Y					Y	Y				
age	-4.5***	-3.99***	-4.43***	-4.67***	see (4)					see (4)					see (3)	see (4)				
	(1.1)	(1.09)	(1.09)	(1.1)																
educ	-4.49***	-5.05***	-4.57***	-4.21**	see (4)					see (4)					see (3)	see (4)				
	(1.2)	(1.2)	(1.2)	(1.21)																
gender	-3.85	-4.38	-3.78	-3.54	see (4)					see (4)					see (3)	see (4)				
	(3.56)	(3.56)	(3.54)	(3.55)																
Controls for tranche and day of week	Y	Y	Y	Y	see (4)					see (4)					see (3)	see (4)				
System Over-identification test							117.86	90	0.026			100.91	90	0.203				87.53	81	0.290

Notes Standard errors are given in parentheses and are corrected for clustering. Standard errors for the LW estimates were calculated by means of a clustered bootstrap with 200 replications.
Significance level: + 10% * 5% ** 1% *** 0.1% (indicated only for OLS results)