JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. **29**: 880–900 (2014)
Published online 9 October 2013 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/jae.2359

CHILD MENTAL HEALTH AND EDUCATIONAL ATTAINMENT: MULTIPLE OBSERVERS AND THE MEASUREMENT ERROR PROBLEM

DAVID JOHNSTON[a], CAROL PROPPER[b,c], STEPHEN PUDNEY[d]* AND MICHAEL SHIELDS[a]

[a] *Centre for Health Economics, Monash University, Melbourne, Australia*
[b] *Business School, Imperial College, London, UK*
[c] *CMPO, University of Bristol, UK*
[d] *ISER, University of Essex, Colchester, UK*

SUMMARY

We examine the effect of survey measurement error on the empirical relationship between child mental health and personal and family characteristics, and between child mental health and educational progress. Our contribution is to use unique UK survey data that contain (potentially biased) assessments of each child's mental state from three observers (parent, teacher and child), together with expert (quasi-)diagnoses, using an assumption of optimal diagnostic behaviour to adjust for reporting bias. We use three alternative restrictions to identify the effect of mental disorders on educational progress. Maternal education and mental health, family income and major adverse life events are all significant in explaining child mental health, and child mental health is found to have a large influence on educational progress. Our preferred estimate is that a one-standard-deviation reduction in 'true' latent child mental health leads to a 2- to 5-month loss in educational progress. We also find a strong tendency for observers to understate the problems of older children and adolescents compared to expert diagnosis. Copyright © 2013 John Wiley & Sons, Ltd.

Received 26 October 2011; Revised 19 June 2013

Supporting information may be found in the online version of this article.

1. INTRODUCTION

Childhood has become the focus of a growing body of research in economics concerned with the closely related concepts of children's well-being, mental health and non-cognitive skills. Much of this interest has been sparked by Heckman's model of life cycle human capital accumulation, which contends that, distinct from cognitive ability, a stock of 'non-cognitive skills' is built up by streams of investment over the life course and influences a wide range of life outcomes (Heckman *et al*., 2006; Cunha *et al*., 2010). A strong motivation for this line of research comes from the belief that IQ or cognitive ability is much less malleable than socio-emotional skills, particularly after the age of 10. From a policy perspective, this would suggest that the returns to interventions targeted at non-cognitive skills are potentially much higher than those focused on cognitive outcomes alone. For example, the Perry preschool intervention program in the 1960s did not raise the IQ of participating children in a lasting way, yet they went on to have better adult outcomes than the control group in a variety of dimensions (Heckman *et al*., 2010). The inference that Perry succeeded because of its impact on attention skills or antisocial behaviours, rather than cognitive ability, is one that is supported by evaluations of more recent childhood interventions which tend to show much larger effects on behaviour (of both parents and children) than on cognitive achievement outcomes (Currie, 2009).

* Correspondence to: Stephen Pudney, ISER, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK.
E-mail: spudney@essex.ac.uk

Mental health conditions are much more common in childhood than most physical conditions. It has been estimated that half of all lifetime mental health disorders start by age 14 (Kessler *et al*., 2007), and a growing body of evidence suggests that prevalence is highest among children from low-income backgrounds. While the relationship between non-cognitive skills and medical conceptions of mental health is unclear (even though in practice they are often measured using the same indicators; e.g. Duncan and Magnuson, 2009), whether interpreted as lack of non-cognitive skills or the existence of a mental health problem, a central concern is the impact that these adverse childhood states have on the process of human capital accumulation and the implications for the intergenerational transmission of economic advantage. It has been recognised recently that mental health conditions are potentially an important channel through which parental socio-economic status influences the outcomes of the next generation. For example, Currie and Stabile (2006, 2007) and Currie *et al*. (2010) found significant impacts of hyperactivity on a range of later educational outcomes in US and Canadian longitudinal data and showed the persistence of these effects. Evidence from the medical literature is rather more mixed but also indicates the importance of mental health problems (Duncan and Magnuson, 2009; Breslau *et al*., 2008, 2009).

A key issue in the empirical study of the impact of child mental health on child outcomes is reliability of measurement. Two types of measure are common in the research literature. Clinical diagnoses are used extensively in psychiatric research, but they have several drawbacks: they are often only available for small, endogenously sampled groups of children; they identify relatively extreme and rare cases (affecting somewhere in the region of 5–10% of children); and they are sensitive to differences in diagnostic practice, which may produce surprising differences between apparently similar groups: for example, diagnosed attention deficit and hyperactivity disorder (ADHD) rates in the USA are double those in Canada (Stabile and Currie, 2006). A second type of measure is derived from a 'screener' module which can be completed quickly by parents, teachers or the children themselves, in the context of large-scale sample surveys. These screeners are designed specifically to identify the symptoms of clinical disorders and are often used as a first step in diagnosing suspected cases—a high screening score being suggestive of a recognised disorder, while lower scores reflect the incidence of symptoms among the 'normal' population. Screener modules are often available in surveys that measure associated outcomes and so provide a way of assessing the relationship between early mental health problems and their consequences. Few data sources are available that give both screening and diagnostic-type information for large representative samples.

Whatever type of information is used, measurement error is an important concern, which has received too little attention in the literature on child mental health and its consequences. There is a substantial body of research suggesting that adults' assessments of their physical health are prone to serious measurement error (e.g. Butler *et al*., 1987; Mackenbach *et al*., 1996; Baker *et al*., 2004; Lindeboom and van Doorslaer, 2004; Etilé and Milcent, 2006; Bago d'Uva *et al*., 2007; Jones and Wildman, 2008; Johnston *et al*., 2009), and this problem is likely to be magnified in the case of child mental health. Children may manifest symptoms differently in different settings, perhaps showing deviant behaviour at school but not at home (or vice versa). They may deny or minimise socially undesirable symptoms when asked by parents or teachers. Informants may also have very different thresholds or perceptions of what constitutes abnormal behaviour in children.

The availability of multiple measures is particularly helpful in dealing with measurement error problems, but there is a strong possibility of observer-specific reporting bias. Evidence in the psychology and medical literatures indicate large disagreements between informants in their assessment of children's psychological well-being. For example, in a sample of US children aged 5–10, Brown *et al*. (2006) found that parents failed to detect half of school-aged children considered to be seriously disturbed by their teachers. Youngstrom *et al*. (2003) found that prevalence rates of comorbidity in a clinical sample ranged from 5.4% to 74.1%, depending upon whether ratings from parent, teacher, child or some combination were used to classify the child. Goodman e*t al*. (2000) suggest that parents

DOI: 10.1002/jae

are slightly better at detecting emotional disorders than teachers but that the opposite is true for conduct and hyperactivity disorders, while the self-assessments of children have less explanatory power than parents or teachers. Johnston *et al*. (2013) show, using data from the Survey of Mental Health of Children and Young People in Great Britain, that estimates of the income gradient in childhood mental health are sensitive to who provides the assessment, with the smallest gradients found when using children's own assessment of themselves rather than those of parents and teachers. A clear implication of this limited body of evidence is that measurement error is substantial and unlikely to be the simple random noise which is assumed by the classical errors-in-variables model. If no observer can be assumed to be unbiased, standard methods cannot be used to identify the true mental health process.

In this paper we make three main contributions. First, we exploit data from a remarkable UK survey (see Section 2) that contains assessments of children's mental health from parents, teachers and the children themselves, to demonstrate the existence of significant biases in all three observers. We do this by using additional diagnostic-style assessments from a panel of expert psychiatric assessors, under the assumption that the experts are able to make the best possible use (in a rational expectations sense) of all available information, but with random variations in the threshold of seriousness they use for generating diagnoses. This model of expert behaviour, set out in Section 3, allows us to identify (up to scale) the parameters of a model representing the distribution of 'true' child mental health conditional on personal and family characteristics.

Second, we estimate the effect of mental health on educational progress, which requires us to overcome a second identification problem (discussed in Section 4), arising from the difficulty in distinguishing the indirect effect of influences on mental health from their direct effect on educational attainment. We use alternative identification strategies to provide parallel estimates of the impact of mental health problems on educational progress, relative to an age-specific norm. The orthodox multiple-indicator latent variable model estimated under the standard assumption of an unbiased observer is not consistent when observers may be biased, and we develop an alternative approach which exploits an exclusion restriction derived from the age-referenced structure of our measure of educational progress. This novel method of instrumental variable (IV) construction does not impose the assumption of an unbiased observer.

Third, our empirical findings cast doubt on the robustness of some of the empirical literature on child mental health. In Section 3, we find strong evidence of different biases in the reports from different types of observers of the child (parents, teachers and children). The standard latent variable method of dealing with measurement error suggests an impact of mental disorders on educational progress much larger than that implied by a simple proxy variable regression; we find that the more appropriate IV method gives results much closer to the naive estimate. Unless we are very sure of our assumptions, it is clearly not enough to presume that an estimation method which allows in some way for the existence of response error is necessarily superior to a naive approach.

2. DATA, DEFINITIONS AND DESCRIPTIVE STATISTICS

The data we use come from the 2004 Survey of Mental Health of Children and Young People in Great Britain, commissioned by the Department of Health and Scottish Executive Health Department, and carried out by the Office for National Statistics. Its aim was to provide information about the prevalence of psychiatric problems among people living in Great Britain, with a particular focus on three main categories of mental disorder: conduct disorders, emotional disorders and hyperkinetic disorders. A sample of children aged between 5 and 16 years was randomly drawn using a stratified sample design (by postcode) from the Child Benefit register. At the time of sampling, Child Benefit was essentially a universal entitlement for parents of all children, so the register provides an excellent sampling frame. Information was obtained in 76% (or 7977) of sampled cases, yielding information gathered from the child's primary caregiver (the child's mother in 94% of cases), from the teacher

DOI: 10.1002/jae

and (if aged 11–16) the young person him/herself. Among cooperating families, almost all the parents and most of the children gave full responses, while teacher postal questionnaires were obtained for 78% of the children interviewed. We focus on a subsample of 6806 white children who have information supplied by their mother, and who have non-missing information for key covariates and mental health measures. The reason for this sample restriction was that ethnic minority and paternal respondent cases were too few for reliable inferences to be drawn about ethnic differences. Inclusion of these groups with associated dummy variables as covariates makes no appreciable difference to the main results.

Child mental health is first assessed in the survey with the Strengths and Difficulties Questionnaire (SDQ). The SDQ is a 25-item instrument for assessing social, emotional and behavioural functioning, and has become very widely used as a measure of the mental health of children. The SDQ questions cover positive and negative attributes and respondents answer each with a response 'not true' (0), 'somewhat true' (1), or 'certainly true' (2). Tables A1 and A2 of the online Appendix (supporting information) give a complete list of the SDQ questions relating to conduct disorder, hyperactivity and emotional problems. In our empirical analyses we use parent, child and teacher SDQ scores that have been constructed in the standard way by summing responses. We carry out the analysis using two alternative indicators:

1. *General Mental Health*: sum of the 15 items for conduct, emotional and hyperactivity disorder.
2. *Hyperactivity*: sum of the 5 items for hyperactivity alone.

Each is normalised to a 0–1 scale. Measure (i) is intended as a general assessment of psychological distress, while (ii) focuses exclusively on the hyperactivity component of ADHD, which has been studied extensively in the research literature and found to be particularly important in some studies. These measures have good internal consistency, with high Cronbach α for the general and hyperactivity measures (see Table I), which are in line with external values reported by Smedje *et al*. (1999).

Following the SDQ is the Development and Well-Being Assessment (DAWBA), a structured interview administered to parents and older children. Although it has limitations, the DAWBA has been found to be an effective diagnostic tool, especially for ADHD (Foreman *et al*., 2009). The DAWBA contains a series of sections, with each section exploring a different disorder; examples include social phobia, post-traumatic stress disorder, eating disorder, generalised anxiety and depression. Each disorder section begins with a screening question that determines whether the child has a problem in that domain. If the child passes the screening question and the relevant SDQ score is normal, the remainder of the section is omitted but, if parent or child indicates that there is a problem or the SDQ score is high, detailed information is collected, including a description of the problem in the informant's own words.

Table I. Sample mean scores for psychological disorders and educational attainment

	Cronbach	Sample means		
	α	All children	No diagnosed disorder	Diagnosed disorder
Parent general SDQ score[a]	0.82	0.218	0.194	0.470
Child general SDQ score[a]	0.79	0.288	0.272	0.443
Teacher general SDQ score[a]	0.86	0.167	0.146	0.411
Parent hyperactivity SDQ score[b]	0.78	0.321	0.293	0.615
Child hyperactivity SDQ score[b]	0.71	0.389	0.372	0.556
Teacher hyperactivity SDQ score[b]	0.88	0.270	0.241	0.596
Educational attainment relative to age norm	—	0.034	0.128	−1.007

[a]Sum of SDQ scores for emotional, conduct and hyperactivity disorders, scaled to the unit interval.
[b]SDQ scores for hyperactivity disorder, scaled to the unit interval. High scores indicate high levels of disorder. Parent, child and teacher scores are present for 6806, 2958 and 5038 cases, respectively. 4891 cases have non-missing educational attainment.

DOI: 10.1002/jae

The DAWBA parent and child interviews respectively take around 50 and 30 minutes, respectively, to complete (Goodman *et al*., 2000). A shortened version of the DAWBA was also mailed to the child's teacher. Once all three DAWBA questionnaires were returned, a team of child and adolescent psychiatrists reviewed both the verbatim accounts and the answers to questions about children's symptoms and their resultant distress and social impairment, before assigning diagnoses using ICD-10 criteria. Importantly, no respondent was automatically prioritised.

Table I provides the sample means for the parent, child and teacher SDQ scores for all children, and for the subsets of children who were and were not diagnosed with an ICD-10 mental disorder. The sample means indicate that teachers report the fewest symptoms (0.167) and that children report the most (0.288). Table I also shows that the SDQ scores of children with a diagnosed mental disorder are two to three times larger than the SDQ scores of children without a mental disorder. Estimated kernel densities of parent, child and teacher SDQ scores are presented in Figure 1. They are positively skewed, with most children exhibiting few symptoms and only a small minority exhibiting many.

The final key variable for our analysis is educational attainment. The survey focuses very much on measurement of mental state and a consequence of this is that educational outcomes are not

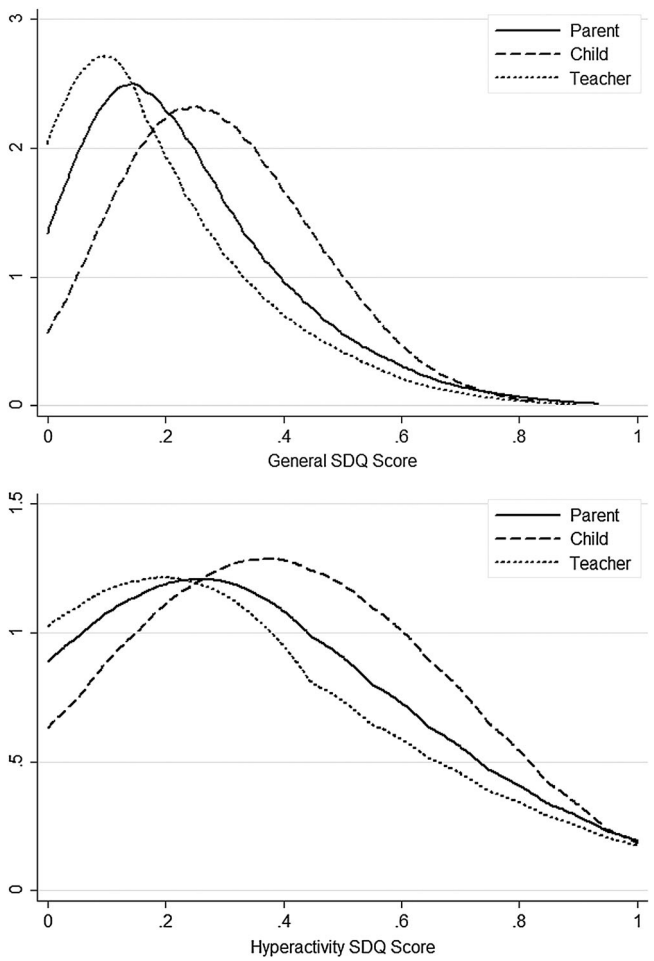

Figure 1. Distributions of SDQ scores for different observers: (a) general mental health; (b) hyperactivity (kernel density estimates: Epanechnikov kernel, bandwidths (a) 0.075, (b) 0.20)

DOI: 10.1002/jae

documented in detail. In particular, the dataset does not contain test score information, and we use instead the one available quantitative measure of general educational progress: the teacher's assessment of the child's scholastic ability relative to other children of the same age. We construct this measure by using teacher responses to the question '*In terms of overall intellectual and scholastic ability*, *roughly what age level is he or she at*?', from which we subtract the child's chronological age. This measure of educational progress is unusual in the economics literature, but the concept of a child's 'mental age' has a long history in child educational psychology—indeed, Intelligence Quotient (IQ) tests are so named because they were originally constructed as the ratio of mental age to chronological age multiplied by 100. The concept also underlies the practice in many educational systems (but not the UK's) of holding children back in a lower grade if he or she has made inadequate progress relative to the norm for that child's age. However, in the UK, the existence of a national school curriculum and associated testing programme means that there is a clear norm of age-specific achievement against which progress can be judged by teachers.

For our sample of children, the average scholastic age gap is 0.034 years, or approximately 2 weeks ahead of actual age (see Table I). The age gap is, however, significantly different from zero for the groups of children with and without mental health problems. For children without a diagnosed mental disorder, the mean gap is 0.128 years, and for those with any disorder the gap is −1.007, implying an average gap between the two groups of around 15 months. Non-parametric estimates of the relationships between parent, child and teacher SDQ scores and educational attainment are shown in Figure 2, which confirms the pattern shown in Table I, but indicates that the relationship is continuous and approximately linear, rather than a discrete distinction between the absence or presence of a disorder (see also Currie and Stabile, 2006). This suggests that identification analysis based on the joint distribution of binary states (see Kreider and Pepper, 2007, for an excellent example) would miss an important feature of the relationship between mental health and education outcomes.

Table A3 of the online Appendix presents sample means for the explanatory covariates used in our analysis. The continuous variables have been scaled to avoid extreme numerical values: age, number of children and log income are divided by 10; and mother's GHQ mental health score is scaled to lie in the [0,1] interval. All other covariates are binary; consequently, the sample means indicate that children with a diagnosed disorder are more likely to be male; live in social housing; have experienced serious adverse life events; and have a parent who is unmarried, less educated, non-employed or with a mental health problem.

3. MODEL-BASED MEASUREMENT OF MENTAL HEALTH

Our model has two components: a model of the complex measurement process for mental health and a relationship between the observed educational outcome and the child's (latent) mental health and other relevant characteristics. The measurement model is based on three main principles. The first is that there exists a 'true' state of mental disorder, S, conceptualised as the (latent) assessment that would be made by experienced psychiatric assessors in possession of fully detailed, multi-source information on the child. This latent measure is the factor which we see as a potential influence on educational development.

Second, we accept that the child's true mental state S is not accurately observable by anyone: not by the parent, the child him/herself, the teacher, the psychiatric assessment team, nor—least of all—by us, the statistical analysts. We assume the SDQ responses from parents, children and teachers are all potentially subject to systematic distortion, which we see as arising either because certain observers (particularly parents and children) may be reluctant to admit the existence of a problem, or may exaggerate minor problems, or because certain aspects of the problem are less visible to certain types of observer.

It is important to realise that any (finitely) biased measure can be reinterpreted as an unbiased measure of a different concept (although not necessarily a theoretically appealing one). Thus any

DOI: 10.1002/jae

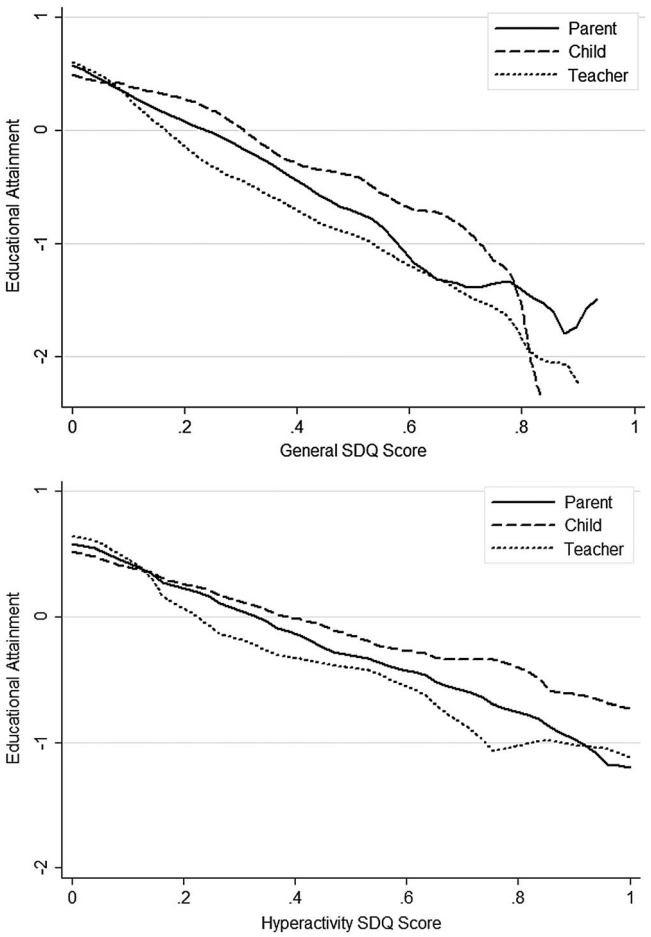

Figure 2. The empirical education-mental health relation: (a) general mental health; (b) hyperactivity (kernel regression estimates: Epanechnikov kernel, rule-of-thumb bandwidth)

single-factor model with biased multiple observers is logically equivalent to a multi-factor model in which each observer measures a different factor. Conti *et al*. (2011) is a recent example of a high-dimensional factor model where the use of observer-specific measures generates additional factors. We would argue that the measurement error approach, involving measurement of a common underlying concept, is a powerful one that has important advantages of parsimony and straightforward theoretical interpretation. It also matches the intention behind the SDQ instrument, which was explicitly designed to achieve comparability across observers (Goodman *et al*., 2000).

The third underlying assumption is that psychiatric assessors make the best use they can of the information available to them, exploiting their experience of diagnosis in a multi-observer setting, where the information reported to them by children and by parents and teachers may be subject to distortions and misinterpretation. Any analysis of measurement error requires an assumption which links some observed measure to the underlying concept that we seek to measure. Our rational expectations assumption for psychiatric assessors provides this link, but how plausible is it? The development of the US Diagnostic and Statistical Manual of Mental Disorders (DSM) and International Statistical Classification of Diseases and Related Health Problems (ICD) provide standardised frameworks for

DOI: 10.1002/jae

diagnosis which help to impose consistency on diagnostic practice and reduce bias of individual assessors relative to the norms set out in the DSM and ICD. Although the psychiatric assessors in our survey do not meet the families, the survey information they have at their disposal is similar to the diagnostic procedures used in connection with the DSM and ICD. There has been a debate in psychology about diagnostic norms and the possibility of bias linked particularly to ethnicity and culture (see USDHHS, 1999), and there is quite strong evidence of a greater readiness to diagnose disorder in black and other minority groups, particularly by white clinicians (Trierweiler *et al*., 2006). DSM-IV and ICD-10 (the versions in force at the time of the survey) both provide for cultural differences to be considered explicitly, but they have been criticised for their Western focus (Kleinman, 1997). Ethnic minorities form a very small proportion of our original survey sample and are not included in the subsample used for our analysis, so the main area of concern over biased psychiatric assessment is not relevant here.

We implement our approach through a latent variable structure with switching between observational regimes, to reflect the different information sets that may be available to psychiatric assessors under different circumstances. The sample consists of a set of n observed children. Child i's 'true' mental health state is S_i, which is related to the child's characteristics and circumstances through a latent regression:

$$S_i = \mathrm{X}_i\beta + U_i \tag{1}$$

where U_i is $N\left(0, \sigma_u^2\right)$. Since S_i is unobservable, we can normalise β and σ_u^2 arbitrarily to fix the origin and scale of S_i. We observe three SDQ scores reported by the parent, child and teacher, Y_{iP}, Y_{iC} and Y_{iT}, all treated as continuously variable measures. The scores derived from parents', children's and teachers' SDQ responses are potentially biased readings of S_i:

$$Y_{ij} = \lambda_j S_i + \mathrm{X}_i\alpha_j + V_{ij}, \quad j = P, C, T \tag{2}$$

where X_i is a vector of variables, available to all observers, reflecting causal factors including the child's personal characteristics, family and social circumstances and the occurrence of past traumatic events. (V_{iP}, V_{iC}, V_{iT}) are jointly normal conditional on S_i and X_i, with zero means and variance matrix $\sum_{YY}$; $\lambda_j - 1$ represents the sensitivity of the observer to the child's true state and α_j captures any measurement distortions linked to specific characteristics of the child and family circumstances. Consequently, an observer of type j gives generally unbiased reports only if $\lambda_j = 1$ and $\alpha_j = 0$. Note that the inclusion of covariates in measurement models is used frequently in psychology to allow for systematic differences ('bias') in the sensitivity of cognitive ability tests and has been used in this way by Carneiro *et al*. (2003) in the economics literature.

The reduced form of (1)–(2) for observer j is

$$Y_{ij} = \mathrm{X}_i\left[\lambda_j\beta + \alpha_j\right] + \lambda_j U_i + V_{ij} \tag{3}$$

and the coefficients $[\lambda_j\beta + \alpha_j]$ describe the mean relationship between the distorted SDQ scores and the child's observable characteristics.

In addition, parents and children are each asked a direct question about whether they perceive there to be a problem with respect to the specific aspect of mental health, yielding two binary indicators, W_{iP}, W_{iC}. These indicators are important, since they play a role in triggering additional questionnaire content, but they are based on the same underlying opinion as revealed by the SDQ and we assume them to contain no additional information, so that $S_i \perp\!\!\!\perp W_{ij} | Y_{iP}, Y_{iC}, Y_{iT}, X_i$. Our rational expectations model of assessors' behaviour has four components:

DOI: 10.1002/jae

1. *Information*. The basic information which is always available to the psychiatric assessment process is $\mathcal{B}_i = \{Y_{iP}, Y_{iC}, Y_{iT}, W_{iP}, W_{iC}, \mathrm{X}_i\}$. If the parent's SDQ score exceeds a specific threshold ($Y_{iP} \geq K_P$) or the parent reports the child's state to be problematic ($W_{iP} = 1$), then a more detailed set of questions is triggered, generating additional information Ω_{iP}; similarly, if the child perceives there to be a problem or his or her SDQ responses exceed a threshold K_C, further information Ω_{iC} is elicited from him or her. Thus the additional contingent information set available to assessors is

$$\mathcal{C}_i = \begin{cases} \varnothing & \text{if} \quad Y_{iP} < K_P, W_{iP} = 0, Y_{iC} < K_C, W_{iC} = 0 \\ \Omega_{iP} & \text{if} \quad (Y_{iP} \geq K_P \text{ or } W_{iP} = 1), Y_{iC} < K_C, W_{iC} = 0 \\ \Omega_{iC} & \text{if} \quad Y_{iP} < K_P, W_{iP} = 0, (Y_{iC} \geq K_C \text{ or } W_{iC} = 1) \\ \{\Omega_{iP}, \Omega_{iC}\} & \text{if} \quad (Y_{iP} \geq K_P \text{ or } W_{iP} = 1), (Y_{iC} \geq K_C \text{ or } W_{iC} = 1) \end{cases} \tag{4}$$

   As external observers, we observe which of these four observational regimes occurs, but not the content of the information sets Ω_{iP} and Ω_{iC}.[1]

2. *Knowledge*. Psychiatric assessors' knowledge and experience gives them the ability to 'purge' informational signals from parents, children and teachers of their bias. Although we do not claim that assessors think in terms of statistical models, this assumption is equivalent to assuming that they know the values of population parameters like $\beta, \lambda_j, \alpha_j$.
3. *Conditionally unbiased expectations*. Assessors make minimum-variance unbiased predictions of S_i, conditional on the available information $\widetilde{S}_i = E(S_i | \mathcal{B}_i \mathcal{C}_i, X_i)$. Standard properties of the multivariate normal distribution imply that this conditional expectation is

$$\widetilde{S}_i = \begin{cases} \mathrm{X}_i\beta + j = P, C, T \sum_{j=P,C,T} b^j_{SY.X} \left[Y_{ij} - \mu^j_Y(X_i) \right] & \text{if} \quad \mathcal{C}_i = \varnothing \\ \mathrm{X}_i\beta + j = P, C, T \sum_{j=P,C,T} b^j_{SY.CX} \left[Y_{ij} - \mu^j_Y(X_i) \right] + \mathrm{b}_{SC.YX} [\mathcal{C}_i - \mu_{\mathcal{C}}(X_i)] & \text{if} \quad \mathcal{C}_i \neq \varnothing \end{cases} \tag{5}$$

where $\mu^j_Y(\mathrm{X}_i) = \mathrm{X}_i\left(\lambda_j\beta + \alpha_j\right)$ is the conditional mean from the reduced form (3). $b^j_{SY.X}$ is the coefficient of Y_{ij} in a population regression of S_i on Y_{iP}, Y_{iC}, Y_{iT} and X_i, and $b^j_{SY.CX}, b_{SC.YX}$ are the coefficients of Y_{ij} and the contingent information $\mathcal{C}_i$ from an extended regression on Y_{iP}, Y_{iC}, Y_{iT} and $\mathcal{C}_i$. Structure (5) implies that assessors make predictions which are optimal linear combinations of the three observers' information signals, *after purging those signals of their bias components*: consequently, the signal receiving the greatest weight is not necessarily the least biased. The term $b_{SC.YX}[\mathcal{C}_i - \mu_{\mathcal{C}}(X_i)]$ represents the contribution of information available to the assessor but not to the statistical analysis and thus, from the point of view of the external observer, only inflates the residual error in $\widetilde{S}_i$.
4. *Diagnosis*. The observed assessment is a binary quasi-diagnosis D_i, indicating a high predicted level of disorder: $D_i = \mathbb{1}\left(\tilde{S}_i \geq \tau\right)$, where τ is the assessor's decision threshold, assumed distributed

[1] Some, but not all, of the contingent diagnostic information is actually available in the dataset, but is not readily usable in a modelling framework because of its complexity and high dimensionality.

DOI: 10.1002/jae

as $N(\mu_\tau, \sigma_\tau^2)$.[2] As an outside observer, the statistical analyst observes the diagnosis D_i and the basic information $\mathcal{B}_i$. The probability of a diagnosed problem is

$$Pr(D_i = 1|\mathcal{B}_i, X_i) = \Phi\left(\sum_{j=P,C,T}\left[\frac{b^j_{SY.X}}{\sigma_\tau}\right] + X_i\left[\frac{\beta}{\sigma_\tau}\right] - \frac{\mu_\tau}{\sigma_\tau}\right) \tag{6}$$

where $E_{ij} = Y_{ij} - (\lambda_j\beta + \alpha_j)$. If contingent information $\mathcal{C}_i$ is available to the assessment process, the probability of a diagnosed mental health problem conditional on the information available to the analyst is

$$Pr(D_i = 1|\mathcal{B}_i, X_i) = \Phi\left(\frac{\sigma_\tau}{\omega_\mathcal{C}}\left\{\sum_{j=P,C,T}\left[\frac{b^j_{SY.CX}}{\sigma_\tau}\right]E_{ij} + X_i\left[\frac{\beta}{\sigma_\tau}\right] - \frac{\mu_\tau}{\sigma_\tau}\right\}\right) \tag{7}$$

where $\omega^2_\mathcal{C} = \sigma^2_\tau + var\left(\sum_j b^j_{SC.YX}[\mathcal{C}_i - \mu_\mathcal{C}(X_i)]\right)$. Thus, conditional on all the observed information in $\mathcal{B}_i$, we have a probit model for the psychiatric assessment, with regime switches in the coefficients of E_{ij} and X_i and in the normalising variance. However, conditional on $\mathcal{B}_i$, these switches are exogenous, so there is no endogenous selection problem as there would be if we conditioned on X_i but not on the SDQ scores Y_{ij}. Note that, if item non-response makes one or more of the SDQ scores unavailable to us and to the assessors, the forms of (5) and (6) or (7) change to take account of the more limited information available.

What can be identified from this measurement model? Equations (2) and (1) imply the following reduced-form SDQ models:

$$Y_{ij} = X_i(\lambda_j\beta + \alpha_j) + (V_{ij} + \lambda_j U_i) \quad , \qquad j = P, C, T \tag{8}$$

Thus regression analysis of the SDQ scores conditional on X_i identifies the reduced-form coefficient vectors $(\lambda_j\beta + \alpha_j)$, which give the response of the SDQ score from observer $j = P, C, T$ to variation in characteristics X. In the $\mathcal{C}_i = \varnothing$ regime, the probit model (6) identifies $b^j_{SY.X}/\sigma_\tau$ for each $j = P, C, T$ and $\left[\beta - \sum_j b^j_{SY.X}(\lambda_j\beta + \alpha_j)\right]/\sigma_\tau$. Consequently, β/σ_τ can be recovered, so that β is identified up to scale.

Estimates were computed using maximum likelihood estimation of a system comprising (6), (7) and (8), parametrised in terms of $\beta/\sigma_\tau, \mu_\tau/\sigma_\tau$, $\left\{\left(b^j_{SY.X}/\sigma_\tau\right), (\lambda_j\beta + \alpha_j) \ , \ j = P, C, T\right\}$ and $\{(\omega_\mathcal{C}/\sigma_\tau), \left(b^P_{SY.CX}, b^C_{SY.CX}, b^T_{SY.CX}\right)/\omega_\mathcal{C}$, each $\mathcal{C}_i \in \mathbb{C}\}$, where $\mathbb{C}$ is the set of non-empty configurations of contingent information ($\mathbf{\Omega}_{iP}, \mathbf{\Omega}_{iC}$ or $(\mathbf{\Omega}_{iP}, \mathbf{\Omega}_{iC})$). To allow for item or individual non-response in the SDQ for children or teachers, as well as the response-triggered contingent information, we allow for four missing data regimes with the following combinations of SDQ scores observed[3]: (i) Y_{iP}, Y_{iC}, Y_{iT}; (ii) Y_{iP}, Y_{iC}; (iii) Y_{iP}, Y_{iT}; (iv) only Y_{iP}. The structure of the vector $b_{SY.CX}/\sigma_\tau$ varies across these four regimes. The scale factors $\sigma_\tau/\omega_\mathcal{C}$ are parametrised as $\exp(\psi_P v_{iP} + \psi_C v_{iC})$, where v_{ij} is the amount of contingent information supplied

[2] Another plausible way of modelling the assessment is to assume that the assessor constructs the probability that the true level of disorder exceeds some critical threshold, then diagnoses a problem if that probability is large enough to cause concern. Under our assumptions, this two-stage process would lead to the same empirical model with slightly different parametrisation.

[3] A few observations involved other combinations of missingness in the SDQ measures; these observations were discarded.

DOI: 10.1002/jae

by observer j, ranging from $v_{ij} = 0$ for no additional information to $v_{ij} = 3$ for contingent information on all three aspects of conduct, emotional disorder and hyperactivity.

Parameter estimates of the psychiatric assessment model are given in Table II. The estimates of $b^j_{SY.CX}/\sigma_\tau$ indicate that, when available, assessors give greatest weight to teacher's SDQ reports, slightly less to the parental report and considerably less to the child's own self-assessment. This relative weighting is a consequence of the different amounts of noise that remain in the parent, child and teacher signals, after they are purged of bias. Note that teachers' assessments are the most informative, but not necessarily the least biased, since the parameters α_T may be large. Indeed, we report evidence below that estimates based on the assumption of zero bias in teachers' assessments are themselves subject to substantial bias. The ψ-parameters are negative, which is consistent with the theoretical prediction that $\sigma_\tau/\omega_C < 1$ and indicates that additional contingent information has value in clarifying the circumstances which led to the problematic self-assessment.

The estimates of β/σ_τ are shown in Table III. They give the influence of the characteristics X on the child's mental state S, using a normalisation of S which is dictated by the variability of psychiatric assessments. Since σ_τ is unknown, scaling is arbitrary and it is only the significance and relative magnitudes of the coefficients that are meaningful here. Maternal education of any kind has a substantial positive influence on the child's mental health, comparable to major adverse life events including loss of a parent through death or divorce/separation and past experience of serious illness or injury. There is some evidence of inter-generational transmission of mental health problems, since the mother's own GHQ measure of mental (ill-)health is found to have a significantly negative influence on the child's mental state. For example, if the GHQ score were to double from the mean level of 0.3 to 0.6, the predicted impact on the child's mental disorder would be around a third as great as the impact attributable to the absence of maternal educational attainment, or to the death of a friend or serious illness or injury during childhood. Indicators of social disadvantage do not have a large influence: housing type and tenure are statistically insignificant and, although log household income has a significant protective effect on child mental health, a very large income increase of around 170% would be

Table II. Estimated parameters of the psychiatric assessment process

Parameter	General mental health		Hyperactivity	
	Estimate	SE	Estimate	SE
(*i*) Y_P, Y_C, Y_T *observed*				
$b^P_{SY.CX}/\sigma_\tau$	4.253***	(0.900)	1.099*	(0.595)
$b^C_{SY.CX}/\sigma_\tau$	1.471*	(0.843)	0.006	(0.682)
$b^T_{SY.CX}/\sigma_\tau$	6.358***	(0.845)	3.461***	(0.549)
(*ii*) Y_P, Y_C *observed*				
$b^P_{SY.CX}/\sigma_\tau$	9.607**	(3.781)	3.575*	(1.897)
$b^C_{SY.CX}/\sigma_\tau$	−1.135	(2.831)	0.160	(2.050)
(*iii*) Y_P, Y_T *observed*				
$b^P_{SY.CX}/\sigma_\tau$	3.264***	(0.515)	−0.043	(0.368)
$b^T_{SY.CX}/\sigma_\tau$	5.814***	(0.475)	3.429***	(0.340)
(*iv*) Y_P *observed*				
$b^P_{SY.CX}/\sigma_\tau$	3.470	(2.267)	1.973	(2.454)
ψ_P	−0.281***	(0.031)	−0.0422***	(0.030)
ψ_C	−0.177***	(0.047)	−0.219***	(0.047)

Significance:
*10%; **5%; ***1%.

DOI: 10.1002/jae

Table III. Estimated coefficients (β/σ_τ) for latent mental disorder equation

Covariate	General mental health		Hyperactivity	
	Estimate	SE	Estimate	SE
Age	0.276	(0.210)	0.428*	(0.244)
Male	0.171	(0.111)	0.059	(0.132)
No. children	0.296	(0.565)	0.097	(0.691)
Social housing	0.043	(0.152)	0.001	(0.178)
Apartment	–0.236	(0.236)	–0.281	(0.303)
Cohabiting	0.340*	(0.183)	0.349	(0.222)
Single	–0.366	(0.258)	–0.327	(0.305)
Widowed/divorced	0.085	(0.248)	0.126	(0.307)
Mother's GHQ	0.398***	(0.046)	0.380***	(0.053)
Mother employed	–0.134	(0.122)	–0.151	(0.146)
Father employed	–0.196	(0.213)	–0.210	(0.259)
Degree	–0.404*	(0.207)	–0.479*	(0.246)
Vocational	–0.342*	(0.193)	–0.422*	(0.222)
A-levels	–0.191	(0.180)	–0.160	(0.216)
O-levels	–0.523***	(0.141)	–0.540***	(0.160)
ln(income)	–0.356***	(0.043)	–0.403***	(0.049)
Parental split	0.230	(0.145)	0.164	(0.178)
Death in family	0.364	(0.233)	0.457*	(0.268)
Death of friend	0.395**	(0.198)	0.412*	(0.233)
Illness	0.255*	(0.142)	0.324*	(0.170)
Injury	0.441**	(0.199)	0.363	(0.257)
Financial crisis	0.239	(0.147)	0.291*	(0.167)
Police trouble	0.278	(0.216)	0.136	(0.265)

Significance:
*10%; **5%; ***1%.

required to produce an effect comparable to that of maternal education or adverse life events. We find no statistically significant evidence of an effect for the child's age (for general mental health) and gender or for the parents' employment or partnership status, in contrast to the SDQ reduced-form estimates (see Tables A4 and A5 in the online Appendix).

If an observer j is unbiased (and thus has $\alpha_j = 0$), then the reduced-form coefficient vector $\lambda_j\beta + \alpha_j$ is proportional to β/σ_τ. Since both are identified, we can rescale each estimate to have unit Euclidean length and carry out a Wald test of their equality (after dropping one redundant element of the difference vector). These tests give strong rejections for all three observers,[4] so we can clearly reject the hypothesis that any of the observers is an unbiased observer, relative to the judgements made by the psychiatric assessors.

Although the distortion parameters λ_j, α_j are not separately identified, some inferences about the nature of the distortions are possible. If, for some observer j and covariate x_k, the identifiable effect of x_k on true mental health (β_k/σ_τ) and on the SDQ report by observer j ($[\lambda_j\beta_k + \alpha_{jk}]$) are of opposite sign, then α_{jk} must have the opposite sign to β_k. This would imply that misreporting by observer j has the effect of attenuating or even reversing the apparent impact of x_k on mental health. We examine this by conducting tests of the hypothesis $H_0^k : (\beta_k/\sigma_\tau)\left[\lambda_j\beta_k + \alpha_{jk}\right] \geq 0, j = P, C, T$ against the alternative $H_1^{jk} : (\beta_k/\sigma_\tau)\left[\lambda_j\beta_k + \alpha_{jk}\right] < 0$ for each variable x_k in turn, using Chen and Szroeter's (2009) test for multiple inequality restrictions. (Note that this is a very conservative procedure, since sign conflicts between β_k/σ_τ and α_{jk} need not generate a corresponding sign conflict between the identifiable coefficients β_k/σ_τ and $[\lambda_j\beta_k + \alpha_{jk}]$.) The test generates significant results only for age, where the joint non-negativity hypothesis can be rejected clearly for the hyperactivity measure ($P = 0.048$) and more

[4] The test statistics are $\chi^2(22) = 1{,}060.4$ (parent); 1,161.3 (child); 862.8 (teacher) for general mental health, and 1,395.2 (P), 1,437.8 (C) and 2,055.9 (T) for hyperactivity.

DOI: 10.1002/jae

marginally for general mental health ($P = 0.106$). This suggests a tendency for observers to understate the problems of older children and adolescents relative to younger children, by the standards of the fully informed expert psychiatric assessment and is perhaps unsurprising, since the early stages of the process of child development are often the focus of special attention, while the problems of older children and adolescents may be less visible to external observers and possibly under-acknowledged by young people themselves.

4. MENTAL HEALTH AND EDUCATIONAL ATTAINMENT

We now turn to the consequences of biased reporting of the child's mental state for inferences about the causal impact of child mental health on educational development. Educational attainment relative to the child's age is denoted by A_i and is assumed to be related to mental health S_i and other covariates X_i as follows:

$$A_i = \rho S_i + \mathrm{X}_i\delta + \eta_i \tag{9}$$

where η_i is a normally distributed regression residual, which may be correlated with some or all of the SDQ residuals V_{ij}. Our results are based on model (9) with dependent variable A_i defined as the difference between the child's educational age and actual age.[5]

Note two potentially limiting features of our analysis using model (9). First, we assume that mental health has a uni-dimensional impact on education. Identification of ρ is a difficult issue and identification becomes more demanding if the state of mental health S_i is treated as multi- rather than uni-dimensional. Our approach is to use a single variable to represent mental health, with alternative broad- and narrow-scope measures used to assess robustness. Thus we again report on two implementations: one where S_i represents a concept of general mental health corresponding to the overall emotional + conduct + hyperactivity SDQ score; the other representing hyperactivity alone. The close correspondence between the findings for these two specifications suggests that the assumption of uniform impact across dimensions of mental disorder is a reasonable approximation.

A second potential shortcoming is that we have only a single quantitative measure of general educational attainment, which is provided by the teacher and therefore subject to observational error just as the teacher's SDQ reports are. There are three reasons for believing that bias in teachers' assessments of educational progress will be a less serious problem than bias in judgements made about mental health. First, teachers are professionals and thus the argument we used to motivate the assumption of unbiased assessment by psychiatrists carries over to teachers in relation to judgements about educational performance. Second, teachers in the UK operate within a tightly defined national curriculum with associated age-specific achievement norms and a rigorous external school monitoring regime. One can credibly argue that this system reduces the scope for bias and performs much the same function of imposing external validity as the DSM-IV and ICD-10 diagnostic frameworks do for psychiatric practice. Third, the measure A_i is a dependent variable so that any independent random noise in A_i only has the effect of reduce precision by inflating $\mathrm{var}(\eta_i)$ rather than introducing bias. This contrasts with the measurement of mental health states, which are used as explanatory covariates and are therefore vulnerable to classical measurement error as well as biased measurement.

Nevertheless, there remains a possibility that teachers' judgements about educational achievement have some element of bias related to childrens' mental states. There is very little evidence available

[5] There are other teacher-reported outcome measures relating to truancy and special needs status but these are essentially indicators of mental health rather than the accumulation of human capital. Other subjective assessments of reading and maths ability are not explicitly norm-referenced and are thus less easily interpretable. A similar model with the dependent variable re-expressed as a proportion of actual age gave similar results but a considerably worse sample fit and those results are not presented here.

DOI: 10.1002/jae

on this issue, although Burgess and Greaves (2013), in a study focused mainly on ethnicity bias, find some tendency for teachers to under-predict the test scores achieved by children in groups with high rates of mental health disorders. If this is the case, then the parameter ρ identified by any of the methods considered here will be larger in magnitude than the true impact of mental health disorder on human capital formation. However, it will not affect the comparison of results from different modelling approaches. If, as we find below, methods using classical assumptions to adjust for measurement error overestimate the impact of mental disorder relative to more defensible methods, then this conclusion is unaffected by any tendency for teachers to confuse mental health problems with slow educational progress.

Since the mental health variable is unobserved, its scale is arbitrary and the magnitude of ρ cannot be interpreted without an appropriate scale normalisation. The identifiable vector β/σ_τ contains the coefficients relevant to S_i/σ_τ, and the coefficient of this variable in the education equation would be $\rho\sigma_\tau$. This is not a helpful normalisation: we would like to be able to rescale the latent variable S to have unit variance, so that its coefficient can be interpreted as the impact on educational performance of a one-standard-deviation change in the measure of mental (ill-)health. However, $\text{var}(S_i/\sigma_\tau)$ equals $\left[\beta' \text{V}\beta + \sigma_u^2\right]/\sigma_\tau^2$ rather than 1, where V is the variance matrix of X_i. The scale parameters σ_u and σ_τ are unknown and there is no convincing prior information on them. We resolve this by using a range of normalisations based on alternative assumptions about the population R^2 of the relationship $S_i = \text{X}_i\beta + U_i$. Given an assumed R^2, and estimates of β/σ_τ and V, $\kappa = \sqrt{R^2\sigma_\tau^2/\beta'\text{V}\beta}$ is a known constant and the rescaling $S_i^* = \kappa S_i/\sigma_\tau$ implies $\text{var}\left(S_i^*\right) \equiv 1$. The corresponding coefficient in the education equation is $r = \rho\sigma_\tau/\kappa$, which we aim to identify.

The reduced form for educational attainment reveals the identification problem we face:

$$A_i = \text{X}_i(\rho\beta + \delta) + (\eta_i + \rho U_i) \tag{10}$$

Even with β known, ρ cannot be uniquely recovered from knowledge of the reduced-form coefficients $(\rho\beta + \delta)$. We explore three alternative identification strategies for the coefficient ρ. The first uses prior information on the residual covariances to reveal the sign of ρ. The second—which we include only to evaluate the performance of the standard latent variable method in this context of biased reporting—uses an assumption that one observer (either parent, child or teacher) is unbiased (and therefore under our rational expectations assumption, also known to be unbiased by the psychiatric assessors), with reporting error uncorrelated with educational attainment: essentially the classical measurement error assumptions used in standard latent variable modelling. The third approach uses an exclusion restriction on the coefficient vector δ, which we implement in two distinct ways.

4.1. Covariance Restrictions

Residual covariances provide information on ρ and this was exploited by Kan and Pudney (2008) in a study of time use also involving biased repeat observations. Our application differs from that study since we do not assume that a particular observer or mode of observation is unbiased and, consequently point identification is not possible here. Let c_j be the residual covariance $\text{cov}(Y_{ij}, A_i|\text{X}_i)$ and $\sigma_{V_j\eta}$ be the covariance between the random component of the measurement error for observer j and the random component of educational progress. Under our assumptions $c_j = \sigma_{V_j\eta} + \rho\lambda_j\sigma_u^2$, implying $\rho = \left[c_j - \sigma_{V_j\eta}\right]/\left(\lambda_j\sigma_u^2\right)$. If we rule out any negative covariance between the random components of SDQ and educational attainment, then $c_j/\lambda_j\sigma_u^2$ is an upper bound on ρ. For parent and child observers, it seems reasonable to assume no correlation between their error in reporting the child's mental state and the random component of the teacher's educational report, so that $\sigma_{V_j\eta} = 0$ and therefore $\text{sgn}(\rho) =$

DOI: 10.1002/jae

$\text{sgn}(c_j)$, $j = P, C$. A one-sided test of $H_0 : c_j = 0$ against $H_1 : c_j < 0$ then establishes the sign of ρ. The test remains valid (but loses power) if $\sigma_{V_j\eta} \geq 0$. In contrast, for teacher observers, we might expect $\sigma_{V_T\eta} < 0$, since a tendency to underrate a child's educational achievement may accompany a tendency to overrate the child's degree of mental disorder due to confounding factors reflecting the 'quality' of the child–teacher match. If so, $\rho \geq c_j / \left(\lambda_T \sigma_u^2\right)$, leaving the sign of ρ ambiguous. We implement the test by estimating simultaneously the reduced-form equations ((8) and (10)) for the SDQ scores and education variable, and using separate one-sided Lagrange Multiplier tests for the residual covariances between the education equation and each SDQ equation. The results are given in Table IV. All residual correlations are negative and significant; they would also be highly significant against two-sided alternatives if adjusted for multiple comparisons by using Bonferroni corrections. Consequently, we have some evidence that the impact of poor mental health on educational progress is negative. Note that Table IV is consistent with the idea of correlated educational and mental health assessments from teachers, since the (negative) residual correlation is larger in magnitude and more significant for teachers than for parent or child.

4.2. Identification with an Unbiased Observer

The most common approach to estimation of models like (9) consists in using a single SDQ score (usually from the parent) as a proxy for the unobserved S_i, which is equivalent to assuming $\alpha_j = 0$ and $\text{var}(V_{ij})$ for some observer j. Examples include Salm and Schunk (2012) and Bartling *et al.* (2012), who use SDQ as a covariate, and Datta Gupta and Simonsen (2010) and von Hinke Kessler Scholder *et al.* (2013), who use it as a dependent variable. This approach fails to address either the classical measurement error problem or the additional problem of biased reporting by parents, children or teachers. The upper panel of Table V shows the estimates of the mental health education impact that results from using one of the SDQ measures, scaled to have unit standard deviation, as a crude proxy for latent mental disorder (full parameter estimates are given in Table A6 of the online Appendix). The estimates suggest that a one-standard-deviation increase in mental disorder has an average effect of retarding educational development by 3.1–5.7 months or 2.7–6.0 months, respectively, for the general measure of mental health and for hyperactivity alone. Note that this is considerably smaller than the unconditional mean gap of 15 months between those with and without a diagnosed disorder (see Table I).

A more sophisticated approach to the measurement error problem which is common in the social sciences is to use the 'structural equation modelling' (SEM) framework, combining a set of measurement equations (2), a latent health equation (1) and a 'structural equation' for education (9) (see Bollen, 1989, for a review).

A single unbiased observer is sufficient to give identification up to scale of the coefficients β, since the reduced-form coefficients in (3) are proportional to β if $\alpha_j = 0$. But, in this framework, a repeat

Table IV. Tests of zero residual covariances between SDQ scores and school performance

	Parent	Child	Teacher
General mental health			
Residual correlation	–0.248	–0.176	–0.332
One-sided *t*-statistic[a]	–17.32	–7.97	–22.96
Hyperactivity			
Residual correlation	–0.273	–0.156	–0.343
One-sided *t*-statistic[a]	–19.10	–7.06	–23.71

[a]Computed as correlation $\times \sqrt{n}$.

DOI: 10.1002/jae

Table V. The estimated mental health-education effect: unbiased observer

	General mental health			Hyperactivity		
	ρ x SD(S_i)	SE	R^2	ρ x SD(S_i)	SE	R^2
SDQ proxy	*Least-squares regression with SDQ proxy*					
Parent	–0.367***	(0.021)	0.172	–0.395***	(0.020)	0.184
Child	–0.258***	(0.032)	0.169	–0.224***	(0.032)	0.163
Teacher	–0.472***	(0.020)	0.214	–0.497***	(0.020)	0.221
Respondent assumed unbiased	*Latent factor model with unbiased observer*					
Parent	–0.718***	(0.031)	0.320	–0.704***	(0.030)	0.263
Child	–0.660***	(0.034)	0.195	–0.683***	(0.036)	0.216
Teacher	–0.676***	(0.032)	0.233	–0.708***	(0.032)	0.271

Note: Standard errors in parentheses; significance:
*10%; **5%; ***1%. All models include the covariates listed in Table II.

observation is required to identify ρ in addition to β. One possibility is to assume that another of the non-professional observers is also unbiased, but a more credible strategy is to retain the assumption of unbiased psychiatric assessments, so that we have two unbiased measures. The further assumptions required for identification are that the SDQ measurement error is independent of the true mental state and educational attainment: $V_{ij}U_i, \eta_i$ for a specific observer $j \in \{P,C,T\}$. This gives three sets of estimates as we take each observer in turn to be the one who is unbiased. Note that ρ is fully identifiable here, but we report it, in the lower panel of Table V, in the normalised form $\rho\sqrt{\beta' \mathrm{V}\beta + \sigma_u^2}$, representing the effect on the mean educational deficit of a one-standard-deviation increase in latent mental disorder. We are also able to infer and report the value of R^2 in the latent mental health equation. These estimates would suggest a substantial causal effect in the range 7.9–8.6 months' educational deficit for a one-standard-deviation increase, using either the general mental health or hyperactivity measure, and an R^2 of around 0.2–0.3 for the latent mental health equation which, as one would expect, exceeds the R^2 statistics for the SDQ proxy regressions, which are depressed by the measurement noise they contain.[6]

4.3. Exclusion Restrictions on δ

We now dispense again with the assumption of unbiased observation and consider exclusion restrictions as a source of identification. Define b to be the reduced-form coefficient vector, $\rho\beta + \delta$, for educational performance. A zero restriction on the kth coefficient in δ implies that the corresponding coefficient in b is $\rho\beta_k = (\rho\sigma_\tau)(\beta_k/\sigma_\tau)$ and, since β/σ_τ is identified from the measurement model, the coefficient $(\rho\sigma_\tau)$ relevant to this normalisation is identified uniquely as the ratio of the kth elements of b and β/σ_τ. The coefficient $\rho\sigma_\tau$ can then be rescaled in the form $r = \rho\sigma_\tau/\kappa$, which is interpretable as the impact of a one-standard-deviation change in mental health. The main problem with this approach is finding exclusion restrictions which can be strongly justified a priori—there are few factors influencing mental health which can confidently be asserted to have no direct causal influence on educational attainment.

[6] Simpler IV methods are also possible and, like the SEM approach, give large estimated impacts. For example, using the parental SDQ score for general mental health as the (presumed) unbiased measure, with the teacher SDQ as instrument, the estimate of ρ is -1.3; using both child and teacher SDQ scores as instruments gives $\hat{\rho} = -0.99$ but a significant Sargan test statistic of 33.96, which is consistent with the presence of bias in the parental SDQ score.

DOI: 10.1002/jae

Table VI. The estimated mental health-education effect: exclusion restrictions

	General mental health			Hyperactivity		
	R^2=0.1	R^2=0.25	R^2=0.4	R^2=0.1	R^2=0.25	R^2=0.4
Loss of friend						
Scaled estimate	–0.129	–0.082	–0.064	–0.137	–0.087	–0.069
SE	(0.116)	(0.073)	(0.058)	(0.133)	(0.084)	(0.066)
Age						
Scaled estimate	–0.383***	–0.242***	–0.191***	–0.332***	–0.210***	–0.166***
SE	(0.134)	(0.085)	(0.067)	(0.117)	(0.074)	(0.059)

Only one of the covariates X_i is a plausible candidate for a direct zero restriction on δ. Some 6.8% of sampled children are reported by the mother to have experienced the death of a friend and the reduced-form coefficients confirm that these events have an impact on SDQ scores (Tables A4–A5 of the online Appendix). Unlike the loss of a parent (which may change the resources of parental time and resources invested in the child's education), or injury or illness experienced by the child him/herself (which may interrupt schooling and study time), it is reasonable to argue that the loss of a friend has no direct impact on the child's education, but only an indirect one through his or her mental state. Two concerns have been raised about the exclusion of this variable: its high prevalence rate which might indicate response bias; and the possibility of correlation with socially graded unobserved factors such as neighbourhood deprivation. Evidence on the first of these is very sparse, but Fletcher *et al*. (2013) report an 8% prevalence rate in the USA for the death of a sibling by age 25. Although US child death rates are higher than those in the UK,[7] the network size suggested by most surveys of children's friendship relations is typically about 5 or 6 (see Conti *et al*., 2013, for example), which far exceeds the number of children per family with children (slightly under 2 for both the USA and UK). Consequently, the survey prevalence rate of 6.8% is broadly consistent with external evidence. There remains a possibility that child mortality may act as a proxy for unobserved factors, such as neighbourhood deprivation, imparting an upward bias to our estimate of ρ. While this cannot be settled definitively, there is some available evidence. Using data from UK neighbourhood statistics for 2010, we find that a regression of the mortality rate in the 5–14 age group on the official index of multiple deprivation gives an R^2 of 0.024 for males and 0.010 for females. Using our survey data and regressing the death of a friend variable on observed family characteristics likely to be associated with neighbourhood deprivation (social tenancy, log income and degree-level education) gives an R^2 ranging from 0.0006 to 0.0085; the overall multiple R^2 is 0.01. These figures suggest only a modest social gradient in child mortality and thus limited scope for bias.

The estimates produced by imposing this exclusion are presented in Table VI, scaled to correspond to R^2 levels in the range 0.1–0.4 for the latent mental health equation. Although the standard errors are larger than we would like, so that the estimated impact is not significantly different from zero, it is still possible to reject unambiguously the hypothesis of an 8- to 9-month impact for a one-standard-deviation increase in mental disorder, as suggested by the conventional latent factor analysis.

As an alternative to this direct a priori restriction, we also exploit a restriction on the effect of age which is suggested by the age-referenced nature of our educational attainment variable, A_i, derived from teachers' assessments of the child's educational age. Let e_i, a_i and X_i represent, respectively, the absolute level of the child's achievement, his or her age, and other personal characteristics, and write the age-specific achievement norm used by teachers as $T(a)$, so that the child's educational age

[7] Fletcher *et al*., 2013, report a mortality rate of 59 per 100,000 for the 1–14 age group, while data for England and Wales suggest an average rate for the years relevant to our sample cohorts of around 17 per 100,000.

DOI: 10.1002/jae

reported by the teacher is $T^{-1}(e_i)$. Now make the further assumptions that: (i) teachers use the population average as their norm, so that $T(a) = E(e|a)$; and (ii) achievement is generated by a normal regression structure: $e\big|a, \mathrm{X} \sim N\left(\theta_1 a + \theta_2 \mathrm{X}, \omega_e^2\right)$. Then our education variable is $A_i = T^{-1}(e_i) - a_i = [e_i - \theta_2 E(\mathrm{X}_i|a_i)]/\theta_1 - a_i$ and its conditional distribution is

$$A_i \mid a_i, \mathrm{X}_i \sim N\left((\theta_2/\theta_1)\left[\mathrm{X}_i - E\left(\mathrm{X}_i\big|a_i\right)\right],\ \omega_e^2/\theta_1^2\right) \tag{11}$$

This implies that A_i is independent of age if the covariates X_i are measured from age-specific means, implying an exclusion restriction on the education equation. The strong age gradient in the onset of mental health disorders documented in the psychology literature (Kessler *et al*., 2007) gives this restriction-identifying power. The sample is large enough to permit the removal of age-specific means to be done non-parametrically, rather than modelling the relationship between X and age explicitly.

The lower panel of Table VI gives the results from exploiting the age-referenced nature of the education variable in this way. It shows that the classical measurement error analysis based on the assumption of an unbiased parent, child or teacher observer exaggerates the causal impact of mental health problems on the development of human capital through schooling. While the unbiased observer approach suggests that a one-standard-deviation increase in mental disorder causes on average an 8- to 9-month delay in educational development, the age restriction indicates a much smaller effect of around 2–5 months. Again, there is no evidence of any difference between the impact of general mental health (covering hyperactivity, emotional and conduct disorders), or hyperactivity alone.

These estimates based on exclusion restrictions both suggest a considerably smaller impact of mental health on educational progress than would be suggested by methods based on the assumption of an unbiased parent, child or teacher observer of the child's mental state. The age exclusion, in particular, is a natural assumption to make, since it exploits the logical structure of our particular measure of educational attainment to generate an identifying restriction. It is striking that the estimated impact that results is broadly similar to the result obtained using SDQ variables as crude proxy variables (Table V), while the more sophisticated latent variable model with an unbiased observer produces considerably larger estimates. Of course, there is no necessity for this to be a general result, but it underlines the proposition that, outside the unrealistic world of classical measurement error, the consequences of dealing with partial and error-prone observations can produce results that differ greatly from the simple reversal of attenuation bias.

5. CONCLUSIONS

We have focused on the role of child mental health as an influence on educational attainment, addressing a set of problems related to the measurement of the child's state of mental health. These measurement difficulties generate two distinct identification problems. The first relates to estimation of the relationship between mental health and personal and family characteristics: the strong evidence of bias in the reports given by parents, children and teachers means that the classical conditions for irrelevance of measurement error in a regression dependent variable are not met. We have overcome this by using a unique dataset which includes a detailed psychiatric assessment, together with a theory (essentially rational expectations) of the behaviour of these assessors, to identify a latent mental health model. However, a second identification problem arises when the educational process is introduced, since natural measures of mental health generated from this latent model are collinear with other explanatory covariates used in the education model. We use two alternative exclusion restrictions which can be argued to be valid theoretically and have sufficient empirical power to contribute useful identifying information. One is the experience of a death of a childhood friend, which is hypothesised to influence education only indirectly through its impact on the mental health of the child. The second is an age restriction which flows from the age-referenced nature of our educational attainment measure.

DOI: 10.1002/jae

We have found that mental disorders are strongly influenced by family history and background, particularly by the mother's own mental health and education, and also by major adverse life events such as the death of a friend or serious illness or injury. The decision-making by expert assessors, which is the key to these conclusions, places greatest weight on the views of teachers, rather less on those of parents and little weight on the self-assessments by young people themselves. Diagnostic behaviour by psychiatric assessors reflects the configuration of information that is available to them and adjusts for the biases inherent in different types of observer.

The impact of mental disorder on educational attainment is significant and, using our preferred strategy based on exclusion restrictions, appears to be important—a loss of approximately 2–5 months educational progress for a one-standard deviation increase in 'true' latent mental disorder. This is closer to the estimate generated by a crude proxy-variable regression which ignores the measurement error problem, than the much larger estimate produced by a multi-indicator latent variable model based on the assumption that at least one of the non-expert observers is unbiased.

On a methodological level, this study exemplifies four important points. First, the measurement error in survey reports of children's mental state is large, not uniform across types of observer (parents, children and teachers) and far from the 'classical' measurement error assumptions embodied in standard latent factor models. The biases that result from the sort of measurement difficulty addressed in this paper can be complex and unexpected in structure and direction. Making allowance for this non-standard form of observation error makes a substantial difference to research findings on issues like the socio-economic gradient in child mental health.

Second, like many other important research issues in the social sciences, the link between child mental health and educational attainment is beset by identification difficulties, and the preferred strategy of using controlled (or 'natural' quasi-)experiments is unavailable because of the nature of the phenomena of interest. Despite this, it has been possible to draw some important conclusions.

Third, this application shows that an attempt to address a measurement error problem inappropriately may make things worse rather than better. In this case, our preferred estimates of the impact of mental disorder on educational progress (which exploit credible a priori restrictions and the specific structure of our measure of educational achievement) are considerably smaller than the range of estimates produced by a conventional latent variable analysis based on the assumption of an unbiased observer—and are much closer to estimates from crude proxy variable regressions. If we are interested primarily in the mental health–education effect, the extra sophistication of the latent variable approach would be detrimental. One cannot, of course, rely on the superiority of naive estimates as a general proposition, but it is important to look carefully at the assumptions underlying more sophisticated approaches.

Finally, we have shown the value of evidence that combines standard survey self-reported information with deeper expert assessments, bringing us closer to the ideal situation where there exists an unbiased observer. The UK Survey of the Mental Health of Children and Young People provides a model for this sort of evidence and its potential is substantial, particularly if the design could be extended to give a longitudinal picture of the evolution of mental health and human capital accumulation over time.

ACKNOWLEDGEMENTS

We are grateful to four anonymous referees for their comments, and to participants at the 2010 Melbourne Workshop in Mental Health and Wellbeing and the 2011 CeMMaP workshop in Survey Measurement and Measurement Error for helpful discussion. Katharina Janke gave valuable assistance. Johnston and Shields would like to thank the Australian Research Council for funding. Pudney's involvement was supported by the European Research Council (project no. 269874 [DEVHEALTH]), with additional support from the ESRC Research Centre on Micro-Social Change (award no. RES-518-285-001) and a Faculty Visiting Scholarship in the Department of Economics and Melbourne Institute at the University of Melbourne.

DOI: 10.1002/jae

REFERENCES

Bago d'Uva T, van Doorslaer E, Lindeboom M, O'Donnell O. 2007. Does reporting heterogeneity bias the measurement of health disparities? *Health Economics* **17**: 351–375.

Baker M, Stabile M, Deri C. 2004. What do self-reported objective measures of health measure? *Journal of Human Resources* **39**: 1067–1093.

Bartling B, Fehr E, Schunk D. 2012. Health effects on children's willingness to compete. *Experimental Economics* **15**: 58–70.

Bollen KA. 1989. Structural Equations with Latent Variables. Wiley: New York.

Breslau J, Lane M, Sampson N, Kessler RC. 2008. Mental disorders and subsequent educational attainment in a US national sample. *Journal of Psychiatric Research* **42**: 708–716.

Breslau J, Miller E, Breslau N, Bohnert K, Lucia V, Schweitzer J. 2009. The impact of early behavior disturbances on academic achievement in high school. *Pediatrics* **123**: 1472–1476.

Brown JD, Wissow LS, Gadomski A. 2006. Parent and teacher mental health ratings of children using primary care services: inter-rater agreement and implications for mental health screening. *Ambulatory Pediatrics* **6**: 347–351.

Burgess S, Greaves E. 2013. Test scores, subjective assessment and stereotyping of ethnic minorities. *Journal of Labor Economics* **31**: 535–576.

Butler J, Burkhauser RV, Mitchell JM, Pincus TP. 1987. Measurement error in self-reported health variables. *Review of Economics and Statistics* **69**: 644–650.

Carneiro P, Hansen KT, Heckman JJ. 2003. Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college choice. *International Economic Review* **44**: 361–422.

Chen L-Y, Szroeter J 2009. Hypothesis testing of multiple inequalities: the method of constraint chaining. CeMMaP Working Paper CWP13/09, Centre for Microdata Methods and Practice.

Conti G, Heckman JJ, Lopes H, Piatek R. 2011. Constructing economically justified aggregates: an application to the early origins of health. In *2nd Annual Health Econometrics Workshop*, Ross School of Business, University of Michigan, Ann Arbor, MI.

Conti G, Galeotti A, Mueller G, Pudney SE. 2013. Popularity. *Journal of Human Resources* (forthcoming).

Cunha F, Heckman JJ, Schennach SM. 2010. Estimating the technology of cognitive and noncognitive skill formation. *Econometrica* **78**: 883–931.

Currie J. 2009. Healthy, wealthy, and wise? Socioeconomic status, poor health in childhood, and human capital development. *Journal of Economic Literature* **47**: 87–122.

Currie J, Stabile M. 2006. Child mental health and human capital accumulation: the case of ADHD. *Journal of Health Economics* **25**: 1094–1118.

Currie J, Stabile M. 2007. Mental health in childhood and human capital. NBER Working Paper 13217.

Currie J, Stabile M, Manivong P, Roos LL. 2010. Child health and young adult outcomes. *Journal of Human Resources* **45**: 517–548.

Datta Gupta N, Simonsen M. 2010. Non-cognitive child outcomes and universal high-quality child care. *Journal of Public Economics* **94**: 30–43.

Duncan G, Magnuson K. 2009. The nature and impact of early achievement skills, attention and behavior problems. In *Rethinking the Role of Neighborhoods and Families on Schools and School Outcomes for American Children*,19–20 November 2009.

Etilé F, Milcent C. 2006. Income-related reporting heterogeneity in self-assessed health: evidence from France. *Health Economics* **15**: 965–981.

Fletcher J, Mailick M, Song J, Wolfe B. 2013. A sibling death in the family: common and consequential. *Demography* **50**:803-826.

Foreman D, Morton S, Ford T. 2009. Exploring the clinical utility of the Development and Well-Being Assessment (DAWBA) in the detection of hyperkinetic disorders and associated diagnoses in clinical practice. *Journal of Child Psychology and Psychiatry* **50**: 460–470.

Goodman R, Ford T, Simmons H, Gatward R, Meltzer H. 2000. Using the Strengths and Difficulties Questionnaire (SDQ) to screen for child psychiatric disorders in a community sample. *British Journal of Psychiatry* **177**: 534–539.

Heckman JJ, Stixrud J, Urzua, S. 2006. The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics* **24**: 411–482.

Heckman JJ, Moon SH, Pinto R, Savelyev P, Yavitz A. 2010. Analyzing social experiments as implemented a reexamination of the evidence from the HighScope Perry Preschool Program. *Quantitative Economics* **1**: 1–46.

Johnston DW, Propper C, Shields MA. 2009. Comparing subjective and objective measures of health: evidence from hypertension for the income/health gradient. *Journal of Health Economics* **28**: 540–552.

DOI: 10.1002/jae

Johnston DW, Propper C, Pudney SE, Shields MA. 2013. The income gradient in childhood mental health: all in the eye of the beholder? *Journal of the Royal Statistical Society, Series A* (forthcoming).
Jones AM, Wildman J. 2008. Health, income and relative deprivation: evidence from the BHPS. *Journal of Health Economics* **27**: 308–324.
Kan M-Y, Pudney SE. 2008. Measurement error in stylized and diary data on time use. *Sociological Methodology* **38**: 101–132.
Kessler RC, Amminger GP, Aguilar-Gaxiola S, Alonso J, Lee S, Ustun TB. 2007. Age of onset of mental disorders a review of recent literature. *Current Opinions in Psychiatry* **20**: 359–364.
Kleinman A 1997. Triumph or pyrrhic victory? The inclusion of culture in DSM-IV. *Harvard Review of Psychiatry* **4**: 343–344.
Kreider B, Pepper JV. 2007. Disability and employment: reevaluating the evidence in the light of reporting errors. *Journal of the American Statistical Association* **102**: 432–441.
Lindeboom M, van Doorslaer E. 2004. Cut point shifts and index shift in self-reported health. *Journal of Health Economics* **23**: 1083–1099.
Mackenbach JP, Looman CWN, van der Meer JBW. 1996. Differences in the misreporting of chronic conditions, by level of education: The effect of inequalities in prevalence rates. *American Journal of Public Health* **86**: 706–711.
Salm M, Schunk D. 2012. The relationship between child health, developmental gaps, and parental education: evidence from administrative data. *Journal of the European Economic Association* **10**: 1425–1449.
Smedje H, Broman JE, Hetta J, von Knorring AL. 1999. Psychometric properties of a Swedish version of the 'Strengths and Difficulties Questionnaire'. *European Child and Adolescent Psychiatry* **8**: 63–70.
Trierweiler SJ, Neighbors HW, Munday C, Thompson EE, Jackson JS, Binion VJ. 2006. Differences in patterns of symptom attribution in diagnosing schizophrenia between African American and non-African American clinicians. *American Journal of Orthopsychiatry* **76**: 154–160.
USDHHS. 1999. Mental Health: A Report of the Surgeon General. Department of Health and Human Services, National Institute of Mental Health, Rockville, MD.
von Hinke Kessler Scholder S, Smith GD, Lawlor DA, Propper C, Windmeijer F. 2013. Child height, health and human capital: evidence using genetic markers. *European Economic Review* **57**: 1–22.
Youngstrom E, Findling RL, Calabrese JR. 2003. Who are the comorbid adolescents? Agreement between psychiatric diagnosis, youth, parent, and teacher report. *Journal of Abnormal Child Psychology* **31**: 231–245.

DOI: 10.1002/jae