*Annual Review of Economics*

# The Econometrics of Early Childhood Human Capital and Investments

Flavio Cunha,[1,2] Eric Nielsen,[3] and Benjamin Williams[4]

[1] Department of Economics, Rice University, Houston, Texas 77251, USA;
email: Flavio.Cunha@rice.edu

[2] National Bureau of Economic Research, Cambridge, Massachusetts 02138, USA

[3] Division of Research and Statistics, Federal Reserve Board, Washington, DC 20551, USA

[4] Department of Economics, George Washington University, Washington, DC 20052, USA

Annu. Rev. Econ. 2021. 13:487–513

The *Annual Review of Economics* is online at economics.annualreviews.org

https://doi.org/10.1146/annurev-economics-080217-053409

JEL codes: C31, J13

## Keywords

early childhood, human capital, measurement error, cardinality

## Abstract

This article reviews recent developments in the econometrics of early childhood human capital and investments. We start with a discussion about the lack of cardinality in test scores, the reasons it matters for empirical research on human capital, and the approaches researchers have used to address this problem. Next, we discuss how the literature has accounted for the errors in human capital measurements and investments. Then, we focus on the estimation of production functions of human capital. We present two different specifications of the production function and discuss when to use one versus the other. We describe how researchers have addressed cardinality, measurement errors, and endogeneity of inputs to estimate the technology of skill formation. Finally, we take stock of the work to date, and we identify opportunities for new research directions in this field.

# 1. INTRODUCTION

The recent literature in economics pays considerable attention to the human capital formation taking place before children start school. In the last two decades, we have seen remarkable developments in theoretical and empirical work whose goal is to understand the role early human capital formation plays in the subsequent evolution of human capital and the determination of socioeconomic outcomes in adulthood (see summaries of the literature in, e.g., Heckman & Mosso 2014, Almond et al. 2018).

One branch of this literature has tried to understand the constraints families face when investing in children's human capital, such as credit constraints (e.g., Caucutt & Lochner 2020) or informational constraints (e.g., Cunha et al. 2013, Boneva & Rauh 2018, Attanasio et al. 2019). Additionally, a growing literature has started to apply developments in behavioral economics to understand the sources of inequality in investments and human capital development in early childhood (e.g., Kalil 2014).

Researchers have also evaluated several human capital formation programs, such as interventions that target children directly by improving the quality of nonparental care (e.g., the Perry Preschool Program; see Heckman & Karapacula 2019a,b) or indirectly through home-based parenting programs [e.g., the Jamaica Psychosocial Stimulation and Nutritional Supplementation Program, as evaluated by Attanasio et al. (2014), Gertler et al. (2014)]. There is also a vibrant area of research on the impacts of public policies such as cash transfers (e.g., Aizer et al. 2016) or the Women, Infant, and Children's Program (e.g., Chorniy et al. 2020). Finally, the literature in this field has investigated how the environment children experience in early childhood affects the formation of human capital at a later stage of the life cycle (e.g., Currie 2013, Aizer et al. 2018).

The interest in this topic has encouraged the development of new econometric tools that address two critical empirical challenges. First, early childhood measures of human capital and investments typically do not have a cardinal scale for economic outcomes—that is, the difference between any two values is not itself meaningful. At best, as far as economic outcomes are concerned, these measures are ordinal. This point holds for both cognitive (e.g., the Bayley Scale of Infant Development; see Bayley 1969) and socioemotional (e.g., the Child Behavior Checklist; see Achenbach & Rescorla 2001) measures. Noncardinality implies that standard statistical methods may not lead to valid inference—and order-preserving transformations of the human capital measures could reverse highly cited findings in this (and other) literature. The lack of robustness is worrisome, given the critical role cardinal analyses using these measures have played in guiding both academic research and public policy.

In this paper, we discuss two methods researchers can use to address the lack of cardinality in measures of human capital. The first method simply consists in using ordinal methods where possible. The second method, called anchoring, uses the relationship between observed measures of human capital and cardinal outcomes (e.g., earnings) to rescale the measures to cardinal units. We consider each method's advantages and disadvantages, and we discuss when to use one or the other.

Second, measures of human capital and investments have errors. Researchers use such measures as dependent variables (to evaluate the impacts of interventions in early childhood), independent variables (to study the impact of early childhood on outcomes), or simultaneously as both dependent and independent variables (to estimate the production functions of human capital). The ubiquity of measurement error in early childhood data requires methods that account for such errors, or inference may not be valid.

This review summarizes the application of theoretical research on the econometrics of measurement error to early childhood human capital and investment data. Our goal is not to have a complete treatment of the econometrics of measurement error (for a summary of that literature, see Schennach 2016). Instead, we focus our attention on methods that explore information from multiple measures of the same latent variable and discuss how to map such methods to factor models—the workhorses of empirical studies of early childhood human capital.

Next, we shift our attention to the estimation of human capital production functions. We first describe two different specifications of the technology of skill formation. One specification follows work by Ben-Porath (1967), and the other specification relates to research by Cunha & Heckman (2007). We discuss when to use one model versus the other; our presentation attempts to clarify that these two models are not designed to explain the same phenomena but rather are suited to capturing different features of human development. Then, we summarize findings showing how anchoring, measurement error, and the correlation between observed inputs and unobserved error terms affect estimates of the parameters of human capital production functions.

In the last part of the review, we describe new research directions in this field. In terms of methodology, we describe how the literature has refined the identification arguments about human capital production functions. We summarize critical evidence emerging from studies that measure human capital at later stages of the life cycle, which suggests that errors are nonclassical and biased. If the same features hold for early human capital measures, researchers will need to adapt existing methodologies or to develop new ones to address such forms of measurement error.

Recent data collection innovations have focused on using investment and skills measures that are better suited for understanding development in early childhood. First, there is an increasing focus on measuring skills that undergo sensitive periods of development in early childhood. Similarly, research has started to move away from gross measures of investments [e.g., investment data from time-use surveys, consumer expenditure surveys, or scales such as the Home Observation for the Measurement of the Environment (HOME) Inventory] toward measures that capture the intensity and quality of interactions between children and their adults, regardless of the context (e.g., at home or daycare).

Second, we need to assess human capital development using measures that have a cardinal scale according to the analyst's (economic) objective. The construction and use of cardinal scales will allow researchers to use innovations in the econometrics of measurement error without adapting them to ordinal methods or having to anchor standard, noncardinal measures to adult outcomes.

The review consists of four sections beyond the Introduction and Conclusion. In Section 2, we discuss ordinal versus cardinal scales of measurement and describe two methodologies that address the fact that existing measures of human capital do not have a cardinal scale. In Section 3, we summarize the theoretical literature applying the econometrics of measurement error to research questions in the economics of early childhood human capital. In Section 4, we discuss the specification and estimation of production functions of human capital. In Section 5, we provide our assessment of where more research is needed, with illustrations of ongoing work.

## 2. CARDINALITY

### 2.1. Scales of Measurement

Social scientists use psychometric scales widely. Indeed, such scales are central to most empirical work in diverse areas studying human capital. In child development, we use psychometric scales to measure cognitive and noncognitive skills. In education, we employ such scales to track achievement gaps/trends, measure school and teacher quality, and assess policy changes' causal impacts. Their frequent availability in both survey and administrative data, as well as the value for human flourishing of the skills they measure, motivates the use and exploration of these scales. However, the literature pays little attention to the fact that these scales are ordinal, which affects their proper use.

Stevens (1946) classifies numerical variables into four groups in terms of their scale: nominal, ordinal, interval, and ratio.[1] We review this classification first, as it clarifies much of the following discussion.

- Nominal variables are qualitative, with the numerical values used only as names. For example, the fact that zip code 80012 is twice as large as zip code 40006 is of no consequence.
- Ordinal variables allow researchers to rank observations, but differences need not correspond in magnitude to differences in the underlying characteristic of interest. Percentiles are ordinal variables: While we can rank individuals at the 75th, 50th, and 25th percentiles of the income distribution, we cannot conclude that the gap between the 75th and 50th percentiles equals the gap between the 50th and 25th percentiles.
- Interval (or cardinal) variables are those whose attributes we can represent with numbers in such a way that the numerical differences between points on the scale represent equal differences in said attributes. Dates have an interval scale: The number of days between December 1 and December 15 is always the same regardless of the year. However, interval scales have arbitrary zero points—e.g., year zero varies across calendars and cultures. Thus, whereas differences are meaningful for interval scales, ratios are not.
- Ratio variables are cardinal variables with a clear definition of zero. For example, weight is a ratio scale with zero naturally defined as the absence of mass. We can thus meaningfully say that one barbell weighs twice as much as another regardless of whether we measure them in pounds or kilograms.

This classification depends on the scale and its relationship to the object of study. For human capital, a nominal scale would consist of numerical values (0 and 1) attached to categories such as "no college" and "college," and which group is labeled "1" is arbitrary. An ordinal scale would be the percentile rank in a test score distribution, whereas an interval (cardinal) scale would be educational attainment. Finally, labor income is a measure of human capital that has a ratio scale.

## 2.2. Psychometric Scales Are Not Cardinal

Empirical analyses using psychometric scores typically apply statistical methods, such as mean differences, regression, and factor analysis, that implicitly assume the scores have an interval (cardinal) scale. In this section, we argue that the interval scale assumption—i.e., that a given change in scores has a fixed interpretation everywhere—is unlikely to hold in economic applications.[2]

Stated plainly, it is clear that cardinality is a strong assumption. More importantly, though, its plausibility depends on the context in which the scale is used. Recall that an interval scale is one in which equal differences in the scale correspond to equal differences in the measured attribute. Thus, as the object of measurement changes, the cardinality (or not) of the scale changes as well. The fundamental problem, then, is that economists rarely use psychometric data in their intended contexts. Because these scales were not designed to be cardinal measures of economic concepts like human capital, there is no particular reason to expect that they will work as such.[3]

[1] The classification of measurement scales is an active area of research with alternative schemes, including those by Mosteller & Turkey (1977) and Chrisman (1998).

[2] The discussion here focuses on a single psychometric scale with a single population of test takers. However, cardinality is also a concern for research designs using multiple scales or populations of test takers. The widespread practice of combining scores across settings by converting them to standard deviation (SD) units may generate additional cardinality violations. For example, a 1 SD increase in math for a young child might have a very different meaning than a 1 SD increase in math for a teenager.

[3] Stevens (1946) argues that the scales used in psychology are ordinal even in their intended contexts. Lord (1975) shows for item response theory that there exist infinitely many cardinally distinct achievement scales that fit any item-response data equally well.

As an example, consider a test assessing mastery of a twelfth-grade curriculum consisting of 10 topics. Suppose that the test has 10 equally weighted items, each of which covers one topic of the curriculum. This test has an interval scale for a twelfth-grade teacher because each additional score increment corresponds to the same amount of the curriculum mastered.

By contrast, an economist might instead be interested in the kinds of achievement that contribute to college completion. The test may no longer be cardinal with this different objective: Changes in some parts of the score distribution may correspond to larger or smaller changes in the likelihood of completing college. Thus, the economist may want to emphasize the regions of the score distribution where the students at the margin of completing college are likely to be found.

The relationship between cardinality and linearity is worth examining in some depth, as it highlights a key disagreement between our approach and the one generally taken in economics. Consider the following simple economic model of attaining a college degree. Normalize the utility of not attaining a degree to zero: $U_{i,0} = 0$. Suppose that the utility of attaining a degree is $U_{i,1} = \pi_0 + \pi_1 h_i + \epsilon_i$, where $h_i$ is the individual's human capital and $\epsilon_i$ is a preference shock. Then, $i$ will attain a college degree ($D_i = 1$) if and only if $U_{i,1} \geq U_{i,0}$. If $\epsilon_i$ follows a logistic distribution, we obtain

$$\Pr(D_i = 1 | h_i) = \frac{e^{\pi_0 + \pi_1 h_i}}{1 + e^{\pi_0 + \pi_1 h_i}}. \qquad 1.$$

This model predicts a linear relationship between the log of odds ratio and human capital, that is,

$$\ln \frac{\Pr(D_i = 1 | h_i)}{\Pr(D_i = 0 | h_i)} = \pi_0 + \pi_1 h_i. \qquad 2.$$

Suppose we decide to measure human capital using twelfth-grade test scores, which we assume are cardinal. We can then directly test whether Equation 2 holds in the data. If we reject linearity, what should we conclude? There are two possible answers: (*a*) Our model is misspecified, or (*b*) the test scores, our putative measure of human capital, are not cardinal.

Model misspecification is, of course, a possibility. The usual approach to avoid misspecification bias is to rely on nonparametric methods—i.e., one would allow test scores to have a nonlinear relationship with the log odds ratio. We suspect that this is the dominant response in economics.

A second, neglected possibility is that, in this context, the test scores are not cardinal measures of human capital. If the scores are not cardinal, then fixed score differences will not correspond to fixed differences in human capital. Thus, we would not expect a linear relationship between scores and the log odds ratio even if the model is correctly specified. As a more extreme example, suppose that $h_i$ represents the quality of *i*'s neighborhood but the model is unchanged otherwise. Now suppose we measure $h_i$ using zip codes, and we fail to reject $\pi_1 = 0$. Of course, this analysis is absurd: The numbers comprising zip codes contain no information about neighborhood quality, so the estimated relationship is not informative about the model.

How can one adjudicate between these two possibilities—misspecification and noncardinality? One approach that flips the problem on its head is to define the scale as that which makes the relationship between scores and some observable fit a known parametric form. Item response theory (IRT) is an example of such an approach. It is also the approach taken by anchoring methods that relate scores to economic outcomes (see Section 2.5).

Empirically, test scores are frequently related to economic observables nonlinearly, suggesting that they may not be cardinal in economic applications. **Figure 1** illustrates this point by plotting the empirical relationship between math achievement and subsequent high school and college completion. The probability of completing high school increases rapidly for math score

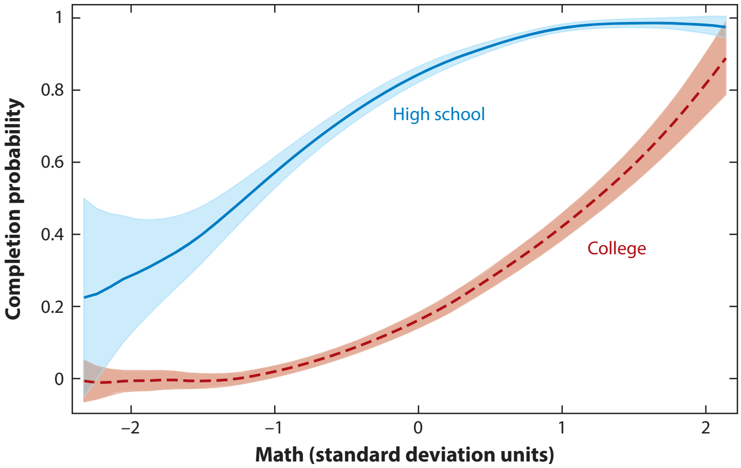

**Figure 1**

School completion conditional on math achievement. The figure shows the local linear polynomial estimated relationships between high school and college completion and the math component score of the Armed Forces Qualifying Test for National Longitudinal Survey of Youth 1979 respondents aged 15–17, standardized by age. Data from Nielsen (2015b).

increases between −2 and 0 standard deviations (SD), consistent with the fact that relatively low-performing students are marginal for high school completion. By contrast, the probability of completing college is fairly constant for these scores but increases rapidly for scores between 0 and 2 SD, consistent with the idea that the marginal students for college completion are from average to above-average in achievement.[4]

The implicit assumption of cardinality shows up in many other contexts as well. Cross-national studies of growth and output often treat indices of corruption or institutional quality cardinally; prominent examples include the studies by Mauro (1995), Hall & Jones (1999), La Porta et al. (1999), and Acemoglu et al. (2001), among many others. Similarly, self-reported well-being measures (i.e., happiness) are commonly treated cardinally, despite ample evidence that doing so is not warranted, as argued by Bond & Lang (2019). In medical research, the efficacy of treatments is often measured not on the primary outcome, such as death, but rather on "surrogate endpoints"—intermediate outcomes that are easier to observe (Aronson 2005).

## 2.3. The Noncardinality of Standard Scales Matters

The argument that psychometric scales are not cardinal is not new.[5] Nonetheless, its importance is still underappreciated in empirical work, where the cardinality assumption is rarely made explicit, much less defended. Stevens (1946, p. 679) exemplifies the attitude toward the use of psychometric scales even among those who recognize the risks of assuming cardinality, arguing that "for this 'illegal' statisticizing there can be invoked a kind of pragmatic sanction: In numerous instances,

[4] Heckman et al. (2006) find similar results for cognitive and noncognitive factors estimated from the same data.

[5] Readers are referred to Stevens (1946), Lord (1975), Cawley et al. (1999), Blanton & Jaccard (2006), Cunha & Heckman (2008), Cunha et al. (2010), Lang (2010), Bond & Lang (2013, 2018), and Jacob & Rothstein (2016), among many others.

it leads to fruitful results." Unfortunately, this is not right: Misapplying cardinal methods generates unknown and perhaps substantial bias, whereby even the sign of the object of interest is not identified except in special cases.

This concern is not just merely theoretical: Recent literature in economics documents the fragility of many standard results to order-preserving transformations of psychometric scales (Bond & Lang 2013, 2018; Nielsen 2015a; Schröder & Yitzhaki 2017). Important empirical questions, such as whether the black–white achievement gap in the United States has decreased over time, hinge on the particular cardinalization chosen for achievement. The robustness of standard estimates to changes in scale turns out to be data and application specific: Some results can be overturned with even mild rescaling, whereas other seemingly similar results are difficult or impossible to overturn.

## 2.4. Ordinal Methods

The most straightforward response to the noncardinality of psychometric scales is to adopt ordinal methods that do not require cardinal data. Where applicable, these methods provide more credible and robust estimates, as they do not rely on possibly questionable cardinality assumptions. Many methods—such as ordinal regression, extensions of Cohen's *d*, and ordinal decomposition—have been proposed and used in economics and education research (e.g., Ho 2009, Ho & Reardon 2012, Ho & Quinn 2020).

One simple ordinal approach builds on the idea that if group *a*'s score distribution first-order dominates group *b*'s, then $\mathbb{E}_a[w(s)] > \mathbb{E}_b[w(s)]$ for any order-preserving transformation $w$. First-order dominance allows one to identify the sign of an achievement difference regardless of the scale's cardinality. In one application demonstrating the power of this approach, Nielsen (2015b) shows that, between 1980 and 1997, the reading achievement gap between high- and low-income youth declined for any cardinalization of achievement. In particular, applying the method for testing first-order dominance developed by Barrett & Donald (2003), the paper shows that low-income achievement improved unambiguously in the sense that the 1997 score distribution first-order dominates the 1980 score distribution. Conversely, high-income achievement unambiguously declined in the same sense.[6]

Analysts can apply ordinal methods to settings other than academic achievement gaps. For instance, other scales of human capital, investment, and noncognitive skills, such as the Home Observation for the Measurement of the Environment (HOME) Inventory, the Rosenberg Self-Esteem Scale, and Rotter's Locus of Control Scale, could be treated ordinally in descriptive and causal analysis. An open question is how far one can take ordinal methods. For example, researchers commonly estimate school and teacher quality using value-added regression models that treat test scores cardinally. Could ordinal versions of value added be implemented?

Though attractive, ordinal methods are not a panacea. By effectively using less information in the scores, ordinal methods may give indeterminate answers in settings where cardinal methods would be unequivocal. More importantly, ordinal methods cannot give any sense of the importance of the skill differences identified. One score distribution might first-order dominate another, but the economic importance of the difference could be trivial. In this sense, the desire in research to identify important quantities stands in tension with the use of ordinal statistics.

[6] This application requires test scores that are ordinally comparable across cohorts, so $s_i > s_j$ implies that *i* has greater achievement than *j* regardless of their cohort. Nielsen (2015b) uses scores specifically constructed to have this property, but this assumption will be untestable for many cross-cohort analyses.

## 2.5. Anchoring

An alternative solution to the noncardinality of standard scales is to estimate relationships between the observed scores and a cardinal anchor outcome to construct a new cardinal scale. Formally, define the anchored achievement of individual $i$ as

$$A_i \equiv \mathbb{E}[Y_i|s_i], \tag{3.}$$

where $Y_i$ is some cardinal outcome (e.g., earnings) and $s_i$ is the noncardinal measure.[7] As $Y$ is cardinal by assumption, the anchored scale will be cardinal as well. Thus, researchers can meaningfully apply standard statistical techniques to the anchored values. In the child development literature, this approach has been widely applied in cases where $s_i$ is an estimated latent factor measuring skills (see Heckman et al. 2006, Cunha & Heckman 2008, Cunha et al. 2010, among others). In education economics, anchoring has been applied to given scale scores (see Cawley et al. 1999, Bond & Lang 2018, among others).

**Figure 1** illustrates the basic idea. The estimates of school completion conditional on math scores comprise anchored scales: Each fixed increment corresponds to the same increment in expected school completion. The notable differences between the high school and college anchored scales clarify the possibility that the same cardinal analysis applied to scales anchored to different outcomes might yield different answers. Although this dependence on the anchor might seem odd to researchers accustomed to using a single psychometric scale in many different settings, it is entirely natural. Cardinality is context dependent, so a scale adapted to one application might yield different answers compared to a scale adapted to another.

Anchoring can also guide the rescaling of psychometric data in cases in which explicitly estimating anchored scales is not feasible. For example, the relationship between achievement and college completion for many different achievement measures looks roughly like the one shown in **Figure 1**. Rescaling using this general shape might thus be preferable to the cardinal use of given scales even when true anchoring is infeasible.

The usefulness and conceptual appeal of anchoring are apparent. Nonetheless, the approach has some downsides. One practical difficulty is that anchoring requires data linking the test scores to a cardinal outcome. Although anchoring to contemporaneous outcomes is possible, longer-run measures such as school completion and earnings will generally be more relevant anchors in economic applications. However, few data sources allow one to link scores to such outcomes. Moreover, anchoring to longer-run outcomes by necessity introduces substantial lags between the assessment date and the completed analysis. Waiting years for an answer is not feasible, particularly for policy-relevant questions.

The anchor relationship defined by $s_i \mapsto \mathbb{E}[Y_i|s_i]$ must be estimated, introducing a number of complications. First, researchers must take a stand on how to model this relationship. Cunha et al. (2010) relate the latent factors linearly to schooling completed by age 19, although they discuss identification in nonlinear settings. By contrast, Cawley et al. (1999) explicitly search for evidence of nonlinearity. Second, the anchor model's estimation creates an additional, quantitatively important layer of measurement error, through both sampling variation and model misspecification. This measurement error is often handled via factor methods in child development, as done by Heckman et al. (2006).

Bond & Lang (2018), by contrast, pursue an instrumental variable approach to handle measurement error in the anchored relationships. We give here a quick overview of their method,

[7] In principle, one could anchor on conditional quantiles or other functions of $Y$. However, the vast majority of applications anchor to the mean.

as it nicely illustrates the key steps needed to carry out an anchoring analysis. They estimate the anchored score for student $i$ with test score $s_{i,t}$ in year (or grade) $t$ fully nonparametrically as the average outcome $Y$ among all students with the same observed score. We have

$$\hat{A}_{i,t} = \frac{1}{N_{s_{i,t}}} \sum_{j:s_{j,t}=s_{i,t}} Y_j, \tag{4.}$$

where $N_{s_{i,t}}$ is the number of students scoring the same as $i$ in period $t$. The anchored scores estimated via Equation 4 are noisy; let $R_t$ denote their reliability. Estimating this reliability is crucial for many empirical applications. For example, any difference in group means will be biased toward zero by exactly this factor in the case that both the true anchored scale and the measurement error are jointly normal, an assumption Bond & Lang (2018) maintain.

To estimate $R_t$, Bond & Lang (2018) note that because $Y_i$ is a noisy measure of $A_{i,t}$, the ordinary least squares (OLS) estimate of $\gamma_t$ in the regression

$$\hat{A}_{i,t} = \kappa_t + \gamma_t Y_i + \xi_{i,t} \tag{5.}$$

will be an attenuated estimate of $R_t$. To solve this attenuation problem, they instrument using the leave-one-out anchored score from the prior year, obtaining

$$\hat{A}^{*}_{i,t-1} = (N_{s_{i,t-1}} - 1)^{-1} \sum_{j \neq i: s_{j,t-1}=s_{i,t-1}} \hat{A}_{j,t-1}. \tag{6.}$$

The leave-one-out construction is so that $Y_i$ does enter into its own instrument.

This example demonstrates that measurement error in anchoring is a broader concept than is traditionally conceived within psychometrics. For instance, anchoring cannot measure transitory (but real) shifts in skills: Gains that fade out will appear as measurement error.

A weakness of current anchoring methods is that they are not causal. An improvement of $\Delta s$ may not cause an improvement of $\Delta Y(s)$ but rather might reflect the uneven influence of unobservables. This concern is related to the classic problem of cultural bias in testing. As an extreme example, a test of yachting knowledge could well have a steep anchored relationship with lifetime earnings. However, this relationship would presumably reflect almost solely confounders, such as family income, rather than any causal effect of yachting knowledge on earnings. It is possible to mitigate this concern in empirical settings by conditioning on relevant student characteristics such as race, gender, and family income when estimating the anchor model. One can also check that the anchor model is similar across different demographic subgroups, a finding that would suggest that the relationship reflects returns to skill more than the effects of confounders. These approaches are crude and not well explored, however. The development of data and methods to allow for causal anchoring analyses would contribute to this literature.

Because the estimated anchor relationships are not structural, they will not generally be invariant to policy. For instance, **Figure 1** suggests a policy designed to improve the achievement of students in the 0–2 SD range might be particularly effective at increasing college completion. However, implementing such a program successfully on a national scale could affect the anchoring function's shape. Alternatively, such a policy might induce so-called teaching to the test, thereby increasing scores without increasing college readiness and flattening the anchoring relationship.

Anchored scales, or indeed any purportedly cardinal scale, might fall short of the cardinal ideal. The estimation of the anchoring function, for instance, introduces both measurement error and specification error as possible sources of noncardinality. Therefore, researchers should assess the robustness of their conclusions to alternative cardinalizations, for example, by reporting results for as many different anchors as are relevant and feasible. Additionally, researchers can assess robustness by entertaining in a controlled way progressively more extreme rescalings and observing the worst-case consequences for the estimates of interest (e.g., Nielsen 2015a).

## 3. MEASUREMENT ERROR

The estimation of human capital production functions requires measures of both human capital and investments. These measures typically suffer from measurement error. If researchers do not account for such errors, the estimates will have a bias, which, in turn, will generate invalid inferences regarding the process of skill formation. Fortunately, the literature has developed econometric methods to leverage multiple measures or other auxiliary data to avoid measurement error bias. The application of these methods requires a deep understanding of the nature of the measurement error.

In this section, we will focus primarily on measurement error in measures of human capital. Measurement error in test scores, and in other measures of human capital, is a problem that has been extensively studied by psychologists. Psychologists use reliability and validity statistics to quantify various sources of measurement error (see, e.g., Thorndike 1951). For example, one constructs a test's split-test reliability by dividing the test into parallel parts and measuring the correlation between scores on the separate parts. The test-retest reliability accounts for additional measurement error sources by comparing scores on parallel tests taken in different environments. For instruments that rely on human assessment, inter-rater reliability measures the agreement between different raters assessing the same behavior or responses. Standard statistics for measuring validity are based on the correlation between the instrument and a more established measure or correlation with relevant economic, educational, or behavioral outcomes. These various statistics are used to form criteria for designing better tests. This design process reduces but does not eliminate measurement error in test scores.

There are many sources of test-score measurement error. One source is item-specific idiosyncrasies due to factors such as the wording of the question. Split-test reliability statistics, such as Cronbach's alpha, are meant to assess the degree of this type of measurement error. The test-taking environment and incentives can also influence test scores. These measurement error sources could be measured by a well-designed test-retest reliability study, though in practice, this is difficult and rare. Even more problematically, individual personality traits can affect the extent to which the test-taking environment impacts the test scores (Almlund et al. 2011). Another source of measurement error is that knowledge of a particular content area may be related to, or require, knowledge of an ancillary content area. For example, a test of mathematical reasoning may also be influenced by reading comprehension.

Economists have long been concerned with errors in economic measures and have developed many econometric methods for addressing the bias such errors can induce.[8] In a standard linear regression model, classical measurement error in a regressor induces attenuation bias in the coefficient estimates, while classical measurement error in the dependent variable does not. The bias can be in either direction in linear models with multiple error-ridden regressors or nonlinear models with a single error-ridden regressor. There are two traditional approaches to avoiding this bias: using external estimates of the signal-to-noise ratio of the measurement to calculate a bias correction, or using a second measurement as an instrumental variable.

Many studies on the econometric analysis of early childhood skill formation have explicitly incorporated measurement error models, finding that accounting for measurement error is essential. Todd & Wolpin (2003) showed that the implicit assumptions about the nature of measurement error in the models commonly used at the time were very restrictive. Cunha & Heckman (2008),

[8] For classical treatments, readers are referred to, e.g., Aigner et al. (1984) and Wansbeek & Meijer (2001). Chen et al. (2011) and Schennach (2016) provide surveys of modern nonlinear models.

building on previous work by Joreskog & Goldberger (1975) and Hansen et al. (2004), among others, estimated a model of skill formation in childhood that explicitly accounts for measurement error in test scores. They found that estimates that do not account for measurement error imply much smaller self-productivity and cross-productivity of skills and smaller investment productivity effects. They found that failing to account for measurement error even produced the opposite sign for the investment effect. Cunha et al. (2010) estimate a nonlinear model of skill formation and find that not accounting for measurement error produces estimates of the various productivity effects that are in some cases stronger and in other cases weaker, with no apparent pattern. Agostinelli & Wiswall (2016), who also estimate a nonlinear model, find that not accounting for measurement error significantly affects the estimated policy effect of income transfers in different directions for different model specifications.

## 3.1. Item Response Theory

Most cognitive ability tests, most commonly used measures of noncognitive traits, and many parental investment scales consist of a series of questions, or items, with categorical responses. Suppose each item $j$ on a test is perfectly discriminating, meaning that individual $i$ answers it correctly if and only if the underlying trait or ability, $\tilde{\theta}_i$, is above some threshold level, $c_j$. If there are many such items on the test, and the thresholds associated with each item vary enough, then we can identify a narrow interval that $\tilde{\theta}_i$ must lie in, but we cannot pinpoint the location of $\tilde{\theta}_i$ within this interval. Moreover, items are generally not perfectly discriminating, so that, due to idiosyncratic factors such as the wording of a question or a lapse in memory, the individual's response to the item is not a deterministic function of $\tilde{\theta}_i$. This stochasticity prevents us from even being able to narrow $\tilde{\theta}_i$ down to a restricted interval with certainty.

Researchers use IRT to model this intra-test measurement error. In a typical IRT model, it is assumed that the item responses $m_{ij}$ are mutually independent conditional on $\tilde{\theta}_i$ and that $m_{ij}$ equals $\mathbf{1}(g_j(\tilde{\theta}_i) \geq \eta_{ij})$, where $g_j$ is a deterministic function and $\eta_{ij}$ is an idiosyncratic error term (see, e.g., Sijtsma & Junker 2006, van der Linden & Hambleton 2013). Under the conditional independence assumption and with a parametric form for $g_j$ and a distributional assumption on $\eta_{ij}$, we can view this as a standard statistical problem of estimating the parameter $\tilde{\theta}_i$ based on a sample of size $J$, the total number of items. However, even if we know the function $g_j$, these IRT scores will be subject to sampling error. As the model is nonlinear, and it is estimated via maximum likelihood or similar methods, the IRT scores are biased estimates of $\tilde{\theta}_i$, but they are consistent as $J \rightarrow \infty$.

In most cases, the function $g_j$ is not known and instead must be estimated as well. This problem is related to the discussion of the lack of cardinality in psychometric scales in Section 2. When $g_j$ is not known, IRT models must impose location and scale restrictions to produce estimates of $\tilde{\theta}_i$. For example, if $g_j(\tilde{\theta}_i) = \mu_j + \beta_j \tilde{\theta}_i$, then the $\tilde{\theta}_i$ cannot be estimated without additional restrictions, because multiplying any set of estimates by 2 and dividing the $\boldsymbol{\beta}_j$ estimates by 2 produces equally valid estimates of the model.[9]

To resolve this problem, we can assume that $\mu_{j^*} = 0$ and $\beta_{j^*} = 1$ for one specified item $j^*$. This normalization is a cardinalization, but it may not be the desired one. Therefore, we view the IRT score as $m_i = m(\tilde{\theta}_i) + \eta_i$, where $\tilde{\theta}_i$ is the latent construct under a preferred cardinalization, $m(\cdot)$ is an unknown monotonic function, and $\eta_i$ is the sampling error induced by intra-test measurement error. We can also view the percent-correct score (sometimes called the raw score) in this way

[9] There is not a unique maximum likelihood estimate because the model is not identified.

(Williams 2019).[10] In both cases, the magnitude of the measurement error is decreasing in the number of items, so that if $J$ is large we have that $m_i \approx m(\tilde{\theta}_i)$.

Thus, in the rest of this section, we ignore the intra-test measurement error. However, intra-test measurement error is not always negligible. Junker et al. (2012), Lockwood & McCaffrey (2014), Schofield (2014), and Williams (2019) have proposed methods for addressing this measurement error. In many cases, it is safe to allow the error term in the factor models below to include intra-test measurement error.

## 3.2. Factor Models

The latent construct $\tilde{\theta}_i$ represents the skill or trait the test measures, along with any other factors that influence responses to a significant proportion of items on the test. It represents a latent variable, conditional on which of the test items are independent of each other. If, for example, the test-taking conditions vary across test takers and these conditions affect performance on the test, then $\tilde{\theta}_i$ should be viewed as a composite of both the measured skill and these test-taking factors. Only item-specific measurement error should be considered negligible when the number of items is large.

In contrast, let $\boldsymbol{\theta}_i$ represent the vector of latent variables of interest. For example, suppose we wish to estimate the model

$$Y_i = \boldsymbol{\beta}_1'\mathbf{X}_i + \boldsymbol{\beta}_2'\boldsymbol{\theta}_i + u_i. \tag{7}$$

Assuming separability, we have

$$\tilde{\theta}_i = h(\boldsymbol{\theta}_i) + \epsilon_i.$$

Suppose the number of items is large so that we can ignore the intra-test measurement error and write $m_i = m(\tilde{\boldsymbol{\theta}}_i)$. If we assume that the functions $m(\cdot)$ and $h(\cdot)$ are both linear, then we have

$$m_i = \gamma_0 + \boldsymbol{\gamma}_1'\boldsymbol{\theta}_i + \epsilon_i.$$

To see the problem caused by the measurement error, $\epsilon_i$, suppose that $X_i$ and $\theta_i$ are scalar, that both are uncorrelated with $u_i$ and $\epsilon_i$, and that $u_i$ and $\epsilon_i$ are uncorrelated. Then, if we use $m_i$ in place of $\theta_i$ to estimate Equation 7, the bias in the OLS estimator is

$$\boldsymbol{\beta}_2 \frac{(\gamma_1^{-1} - 1)(1 - \rho^2) - \sigma^{-1}}{1 - \rho^2 + \sigma^{-1}},$$

where $\rho$ is the correlation between $X_i$ and $\theta_i$ and $\sigma$ is the signal-to-noise ratio, $\sigma = \mathrm{Var}(\theta_i)/\mathrm{Var}(\epsilon_i)$. If we impose the scale restriction $\gamma_1 = 1$, then this is the familiar attenuation bias, $-\boldsymbol{\beta}_2 \frac{1}{1+\sigma(1-\rho^2)}$. Note that the bias of the OLS estimator of $\boldsymbol{\beta}_1$ will also be a function of $\sigma(1 - \rho^2)$; in the $\gamma_1 = 1$ case, this bias is $\boldsymbol{\beta}_2 \frac{\mathrm{Cov}(X_i,\theta_i)}{\mathrm{Var}(X_i)} \frac{1}{1+\sigma(1-\rho^2)}$.

If we have external estimates of $\sigma$ and $\gamma_1$, then we can use them to resolve the measurement error bias in the above regression. To see this, consider again the case where $\gamma_1 = 1$. If $X_i$ and $\theta_i$ are uncorrelated, or if $X_i$ is not included in the regression model, then an unbiased estimate for $\boldsymbol{\beta}_2$ can be obtained by replacing $m_i$ with $\frac{1}{1+\sigma^{-1}}m_i$ in the regression. More generally, the bias in the OLS estimator arises because $\mathrm{Var}(m_i) = \mathrm{Var}(\theta_i) + \mathrm{Var}(\epsilon_i) \neq \mathrm{Var}(\theta_i)$. So an unbiased estimate

[10] The same is true for transformations (e.g., the percentile transformation) of the IRT score or the percent-correct score.

can be constructed by replacing $\mathrm{Var}(m_i)$ with $\mathrm{Var}(\theta_i) = \frac{1}{1+\sigma^{-1}}\mathrm{Var}(m_i)$ in the OLS formula. If an external estimate of $\sigma$ is not available, additional measurements of $\theta_i$ will generally be required.[11]

Suppose then that a total of $K > 1$ measures are available. The latent construct underlying each measure will generally be different. For example, test-taking conditions may vary across tests. Thus, for each $k = 1, \ldots, K$, suppose the $k$th measure contains a sufficient number of items so that we can write $m_{ik} = m_k(\tilde{\theta}_{ik})$, where $\tilde{\theta}_{ik} = h_k(\boldsymbol{\theta}_i) + \epsilon_{ik}$. Suppose that, for each test, the mapping between $\tilde{\theta}_{ik}$ and $m_k(\cdot)$ and the mapping between $\boldsymbol{\theta}_i$ and $\tilde{\theta}_{ik}$ are both linear. Then, we obtain the standard linear factor model

$$m_{ik} = \gamma_{0,k} + \boldsymbol{\gamma}'_{1,k}\boldsymbol{\theta}_i + \epsilon_{i,k}, k = 1, \ldots, K,$$

where the components of the vector $\boldsymbol{\theta}_i$ are known as the factors and the coefficient vectors, $\boldsymbol{\gamma}_{1,k}$, are the factor loadings. This system of equations can also be written in matrix form as

$$\mathbf{m}_i = \boldsymbol{\Gamma}_0 + \Gamma_1\boldsymbol{\theta}_i + \epsilon_i.$$

Here we have defined $\boldsymbol{\theta}_i$ as the latent variable, or variables, of interest. But this interpretation should be evaluated in relation to the assumptions made about the sources of measurement error, as we discuss below. In the remainder of this section, we will primarily focus our attention on this linear factor model, though nonlinear measurement systems have been considered by Cunha et al. (2010), among others.

### 3.2.1. Factor scores.

When multiple measures are available, they can be used to construct factor scores. If $\Gamma_0$ and $\Gamma_1$ were known, then we could construct an unbiased estimate

$$\hat{\boldsymbol{\theta}}_i = \Gamma_1^{\dagger}(\mathbf{m}_i - \boldsymbol{\Gamma}_0) = \boldsymbol{\theta}_i + \Gamma_1^{\dagger}\boldsymbol{\epsilon}_i,$$

where $\Gamma_1^{\dagger} = (\Gamma_1' W \Gamma_1)^{-1}\Gamma_1' W$ for any nonsingular matrix $W$. Choosing $W = \Delta^{-1}$, where $\Delta$ is the variance-covariance matrix of the vector $\boldsymbol{\epsilon}$, minimizes the variance of the error term, $\Gamma_1^{\dagger}\boldsymbol{\epsilon}_i$. Replacing $\boldsymbol{\Gamma}_0$, $\Gamma_1$, and $\Delta$ with first-stage estimates, we can construct a feasible version of this minimum-variance estimate of the factors. This is known as the Bartlett factor score.

The use of these factor scores as regressors in regression analysis will result in measurement error bias for the same reasons described above. While the Bartlett factor score minimizes this bias by minimizing $\mathrm{Var}(\Gamma_1^{-}\boldsymbol{\epsilon}_i)$, the remaining measurement error bias may not be negligible unless the number of measures is very large relative to the dimension of $\boldsymbol{\theta}_i$.

The so-called regression factor scores, or Thurstone factor scores, avoid this bias in a particular case. In the general case, however, these factor scores still lead to bias. One solution to this problem would be to correct the resulting bias, as described above for the general case of measurement error. Heckman et al. (2013) apply this bias correction method, known in the psychometrics literature as Croon's method (Croon 2002). Another solution is to model both the outcome equation and the measurement model jointly.

### 3.2.2. Identification of the factor model.

Construction of the factor scores requires estimates of the parameters $\gamma_{0,k}$ and $\boldsymbol{\gamma}_{1,k}$. The Bartlett factor scores also require an estimate of the error covariance matrix, $\Delta$. In some cases, external estimates of these parameters may be available, but typically they will need to be estimated in the first stage using the same sample of observed measures. In this section, we discuss the necessary conditions to identify and estimate these

[11] If additional restrictions are placed on the error terms, an unbiased estimator based on higher-order moments exists and additional measurements are not needed. This was first established by Reiersol (1950).

parameters. We first describe some standard approaches before addressing some issues that the recent literature on early childhood development has raised.

Suppose, as is typically assumed, that $\epsilon_{ik}$ is uncorrelated with $\boldsymbol{\theta}_i$ for all $k$, and that $\epsilon_{ik}$ and $\epsilon_{ik'}$ are uncorrelated for all $k \neq k'$. The first of these assumptions means that the common factors, $\boldsymbol{\theta}_i$, do not correlate with other variables that affect the measures. The second means that the error terms in a particular measure do not correlate with the error terms in other measures. We will consider relaxing these assumptions later in this section.

First, we can show, as an example, that if $\theta_i$ is a scalar, then three measurements are sufficient for identification under conventional assumptions.[12] We start by normalizing $\gamma_{0,1} = 0$ and $\gamma_{1,1} = 1$. In this sense, the location and the scale of the factor $\theta_i$ are tied to the location and scale of the measure $m_{i,1}$. From these restrictions, we can identify the mean of the factor because we have that $E(m_{i,1}) = E(\theta_i)$. To identify factor loadings and the variance of the factor, consider the following system of covariance moments:

$$\begin{aligned}\mathrm{Cov}(m_{i,1}, m_{i,2}) &= \gamma_{1,2}\sigma_\theta^2,\\ \mathrm{Cov}(m_{i,1}, m_{i,3}) &= \gamma_{1,3}\sigma_\theta^2,\\ \mathrm{Cov}(m_{i,2}, m_{i,3}) &= \gamma_{1,2}\gamma_{1,3}\sigma_\theta^2.\end{aligned}$$

Note that the following ratios identify the loadings and the variance of the latent factor $\ln h_{i,t}$:

$$\begin{aligned}\gamma_{1,2} &= \frac{\mathrm{Cov}(m_{i,2}, m_{i,3})}{\mathrm{Cov}(m_{i,1}, m_{i,3})},\\ \gamma_{1,3} &= \frac{\mathrm{Cov}(m_{i,2}, m_{i,3})}{\mathrm{Cov}(m_{i,1}, m_{i,2})},\\ \sigma_\theta^2 &= \frac{\mathrm{Cov}(m_{i,1}, m_{i,3})\mathrm{Cov}(m_{i,1}, m_{i,2})}{\mathrm{Cov}(m_{i,2}, m_{i,3})}.\end{aligned}$$

We can then use the variances of observed measures to identify the variances of measurement error because we have

$$\mathrm{Var}(m_{i,k}) = \gamma_{1,k}^2\sigma_\theta^2 + \sigma_{\epsilon_k}^2. \qquad 8.$$

Lastly, we can identify the remaining $\gamma_{0,k}$ for each $k$ from the equation $E(m_{i,k}) = \gamma_{0,k} + \gamma_{1,k}E(\theta_i)$.

Moving beyond the one-factor case, the classical treatment of the identification problem, which relies on the orthogonality assumptions discussed above, is due to Anderson & Rubin (1956). Under these orthogonality restrictions, we have that

$$\mathrm{Var}(\mathbf{m}_i) = \Gamma_1\Phi\Gamma_1' + \Delta,$$

where $\Delta = \mathrm{Var}(\boldsymbol{\epsilon}_i)$ is a diagonal matrix. Anderson & Rubin (1956) assume that $\Phi = I_L$, where $L$ is the dimension of $\boldsymbol{\theta}_i$ and $I_L$ is the $L \times L$ identity matrix. This means that the factors are mutually uncorrelated and that the variance of each factor is normalized to equal 1. However, even with this restriction, for any matrix $R$ such that $RR' = I_L$, we can define $\tilde{\Gamma}_1 = \Gamma_1 R$ and $\tilde{\Gamma}_1\tilde{\Gamma}_1' = \Gamma_1\Gamma_1'$. Anderson & Rubin (1956) discuss various additional restrictions to overcome this limitation. Among these, the most commonly used in the economics literature is the restriction requiring $\Gamma_1$ to be a lower triangular matrix.

[12] Bonhomme & Robin (2009) use higher-order moments to reduce this requirement.

In addition to the restrictions above, Anderson & Rubin (1956) assume that $\Gamma_1$ satisfies a row deletion property. The row deletion property assumed is that if any row is removed, then the remaining rows can be rearranged into two different matrices of rank $L$, which implies the requirement that $K \geq 2L + 1$. Under these assumptions, we can identify all of the parameters of the factor model.

If $\Gamma_1$ satisfies the row deletion property and $\Phi$ is positive definite, then $\text{Var}(m_i)$ can be decomposed into the variation due to the common factors, $\Gamma_1 \Phi \Gamma_1'$, and the variation due to the idiosyncratic components, $\Delta$. The additional restrictions requiring $\Phi$ to be equal to $I_L$ and $\Gamma_1$ to be lower triangular then uniquely identify $\Gamma_1$ from $\Gamma_1 \Phi \Gamma_1'$. However, alternative sets of restrictions are sometimes preferable. For example, the assumption that factors are mutually uncorrelated is undesirable in many applications. A common alternative is to assume that the first $L$ rows of $\Gamma_1$ form a diagonal matrix.

Another alternative is to link restrictions across different measurement systems. Estimating the production function of human capital requires multiple human capital measures at different ages. Thus, in Section 4, we use a different measurement system for each age. One can use restrictions across these systems to obtain identification. For example, Agostinelli & Wiswall (2016) suggest leveraging age invariance of the factor-loading matrix. They assume that $\Phi = I_L$ only at the earliest age. At subsequent ages, $L$ rows of the factor-loading matrix are equal to the corresponding rows in earlier ages.[13]

Orthogonality assumptions are often overly restrictive in situations commonly faced in the econometrics of early childhood development. For example, suppose that several of the $K$ measures of human capital are parental assessments of child development at a particular age. Parents may be imperfect, biased assessors, and their bias will influence all such measures.

Williams (2020) provides a framework for understanding when identification is achievable if $\Delta$ is not diagonal but still contains some zeros. One way to illustrate identification is by viewing one set of measures as proxies for $\boldsymbol{\theta}_i$ and the remaining measures, or some of them, as instrumental variables. Take a set $L$ of the total $K$ test scores,

$$\mathbf{m}_i^{(1)} = \boldsymbol{\Gamma}_{0,1} + \Gamma_{1,1}\boldsymbol{\theta}_i + \boldsymbol{\epsilon}_i^{(1)}.$$

Assuming that $\Gamma_{1,1}$ is invertible, solve for $\boldsymbol{\theta}_i$ and plug the resulting expression into the equation for any of the remaining $K - L$ test scores; we obtain

$$\begin{aligned}\mathbf{m}_{ik} &= \gamma_{0,k} + \gamma_{1,k}\Gamma_{1,1}^{-1}\left(\mathrm{m}_i^{(1)} - \boldsymbol{\Gamma}_{0,1} - \boldsymbol{\epsilon}_i^{(1)}\right) + \epsilon_{i,k} \\ &= \tilde{\gamma}_{0,k} + \gamma_{1,k}\Gamma_{1,1}^{-1}\mathbf{m}_i^{(1)} + \tilde{\epsilon}_{i,k},\end{aligned}$$

where we have $\tilde{\gamma}_{0,k} = \gamma_{0,k} - \gamma_{1,k}\Gamma_{1,1}^{-1}\boldsymbol{\Gamma}_{0,1}$ and $\tilde{\epsilon}_{i,k} = \epsilon_{i,k} - \gamma_{1,k}\Gamma_{1,1}^{-1}\boldsymbol{\epsilon}_i^{(1)}$. Since $\mathbf{m}_i^{(1)}$ is correlated with $\tilde{\epsilon}_{i,k}$, OLS estimation of this equation will be biased. However, if there are $L$ test scores among the remaining $K - L - 1$ test scores, then they can be used as instrumental variables to identify $\gamma_{1,k}\Gamma_{1,1}^{-1}$ subject to the usual exogeneity and rank conditions for instrumental variables. If we further impose the normalization that $\Gamma_{1,1} = I_L$, then $\gamma_{1,\,k}$ is identified.

## 3.3. Dedicated Measurements

Often in empirical settings with multiple measures on multiple latent factors, researchers have some a priori knowledge of which factors affect which measures. Measurement $k$ is said to load on

[13] Although $L$ is equal to 1 in their paper, one can extend the argument to the general case.

factor $\ell$ if $\gamma_{1,k,\ell} \neq 0$. The most common type of restriction used is one wherein some measurements load only on a single factor. We refer to these systems as dedicated measures.

When enough dedicated measures are available, they can be used to estimate factor loadings and to construct scores for each latent factor separately. The availability of dedicated measures lends credibility to the analysis because the method is robust. For example, suppose we have access to three math tests and three reading tests in addition to a measure (perhaps another assessment) that requires both verbal and mathematical ability. We can then estimate a math score and a reading score with the respective dedicated measures only, discarding the combined measure. While this method is both robust and interpretable, it sacrifices statistical efficiency.

Economic relationships, or any other source of a priori information, can impose additional restrictions on a factor model's structure, thus aiding in identification. For example, suppose that we observe four test scores over two different periods (two in each period). Let $\theta_{i1}$ denote latent ability in the first period, and let $\theta_{i2}$ denote latent ability in the second period. We assume that test scores only load on the contemporaneous latent ability. We can write the system of equations as

$$m_{i,1} = \gamma_{0,1} + \gamma_{1,1}\theta_{i1} + \epsilon_{i1},$$

$$m_{i,2} = \gamma_{0,2} + \gamma_{1,2}\theta_{i1} + \epsilon_{i2},$$

$$m_{i,3} = \gamma_{0,3} + \gamma_{1,3}\theta_{i2} + \epsilon_{i3},$$

$$m_{i,4} = \gamma_{0,4} + \gamma_{1,4}\theta_{i2} + \epsilon_{i4}.$$

Standard results do not help in showing the identification of the parameters of this model. If we take the scores from the two periods separately, we only have two measurements. If we take the four scores together, we still have four measurements on two factors, whereas the standard result requires at least five measurements. This model is identified with standard normalizations provided that $\mathrm{Cov}(\theta_{i1}, \theta_{i2}) \neq 0$. The intuition is that the measurements from period 1 provide additional variation in $\theta_{i2}$, and vice versa. Williams (2020) shows more generally how overidentifying zero restrictions or reduced rank restrictions can be used to identify the factor model.

## 4. ESTIMATION OF PRODUCTION FUNCTIONS

The identification and estimation of human capital production functions are active research areas in the econometrics of early childhood development. These functions link the output, human capital at the beginning of the next period (e.g., year or developmental stage), to observed inputs, current-period human capital and investments, and unobserved error terms. We start this section by discussing two different specifications of the technology of skill formation, and we summarize the evidence that justifies each of these parameterizations. Then, we describe how research to date has addressed three challenges to estimating such production functions: the lack of cardinality of test scores, the presence of error in the measurement of human capital and investments, and the correlation between observed inputs and unobserved error terms.

### 4.1. The Specifications of the Production Function of Human Capital

In economics, the typical approach is to assume that human capital is a scalar, and that, at each point in their lives, individuals invest in increasing their stock of human capital. The canonical law of motion of human capital in Ben-Porath's (1967) model represents such a process. Let $h_{i,t}$ and $x_{i,t}$ denote, respectively, child $i$'s stock of human capital and investment at period $t$, and let $\eta_{i,t}$

denote unobserved error terms. The Ben-Porath law of motion of human capital is

$$h_{i,t+1} = (1 - \delta)h_{i,t} + f(h_{i,t}, x_{i,t}, \eta_{i,t}). \qquad 9.$$

According to Equation 9, child $i$ starts period $t$ with a given stock of human capital, $h_{i,\,t}$. The production function $f$ then describes how investments, current human capital, and unobservables combine to increase the next-period human capital stock, with the current stock also depreciating by $\delta$ each period. Such models are typically presented in value-added form, in which the production process's output is given by the change in the human capital stock between $t$ and $t + 1$ (for other variants of the value-added specification, see Todd & Wolpin 2007). Researchers in labor economics (e.g., Heckman et al. 1998, Polachek et al. 2016), macroeconomics (e.g., Lucas 1988), and the economics of education (e.g., Rothstein 2010, Chetty et al. 2014), among other fields, explore both linear and nonlinear variants of the Ben-Porath model.

In all of these applications, the assumption is that the form of human capital modeled reflects the accumulation of knowledge, facts, or skills that individuals can acquire throughout their lives. Cattell (1963) classifies this form of human capital as crystallized intelligence. As an example, Cunha & Wolpin (2020) adopt the parameterization in Equation 9 to study the impact and mechanisms of the JumpStart Program, a parenting program implemented by the Alief Independent School District (ISD) in Houston, Texas. The program's goal is enabling parents to help their children learn skills that, according to the Alief ISD, will prepare them for the district's pre-K curriculum. The program trains parents to teach their children the names of colors and to recognize letters, numbers, and shapes, all of which are elements of Cattell's concept of crystallized intelligence.

However, there are elements of human capital whose acquisition is not synonymous with increasing stocks of a particular skill over time, because their development takes place within a specific window of opportunity. Besides, the investments in such skills may spill over into the formation of other skills that individuals acquire in later stages of the life cycle. One illustrative example is the capacity to recognize sounds in a language (phonetic discrimination). When children are born, they can recognize phonetic contrasts that exist in many languages (e.g., Eimas & Miller 1992). When they are 6 months old, infants start to tune their capacity to discriminate phonetic distinctions common in their native language (Werker et al. 1981, Kuhl et al. 1992).

Research has shown that an exogenous manipulation of the language environment, combined with social interaction, influences this tuning process in young children. Kuhl et al. (2003) randomly selected American-born children to either a control group or one of three intervention groups. Researchers exposed the children in the three intervention groups to a 1-hour dose of Mandarin per week for 12 weeks. The groups differed in the means of exposure. In the first group, the children socially interacted with a native speaker of Mandarin. Thus, they not only got exposure to Mandarin but did so with multiple opportunities for joint attention. In the second group, the social interaction was removed, and the children watched a video of the same speaker talking about the same content as in the first group. In the third group, the children heard only the recorded audio of the same content, again with no interaction. The results showed that, at age 12 months, the children in the first group maintained their capacity to recognize frequently used sounds in Mandarin, and their ability to do so was comparable to typical same-aged children living in Taiwan. By contrast, the children in the control group and in the other two treatment groups (video and audio only) did not maintain the capacity to recognize sounds.

Like adults, 1-year-old infants cannot discriminate sounds that are uncommon in their native language. While it might seem that such tuning is a disadvantage, this is not the case. First, if the capacity to recognize sounds consumes brain resources (e.g., memory), maintaining the capacity to recognize uncommon phonetic contrasts in the native language would be wasteful. Second, the tuning process allows the brain to reallocate unused resources to acquire other language skills (Kuhl 2004, Werker & Tees 2005). Indeed, sound recognition is an essential input in the formation of expressive language skills (see Aslin 2014). Therefore, investments in the development of a tuned phonetic recognition system influence more complex language skills in later stages of the life cycle.

The finding by Kuhl et al. (2003) that only the children in the social interaction group were able to maintain the capacity to recognize typical Mandarin phonemes suggests the importance of joint attention as a form of investment in developing certain language skills.[14] A literature overview by de Villiers & de Villiers (2014) forms the basis of a conjecture that language interventions that promote joint attention are likely to influence, in a later stage, the development of theory of mind (Tomasello 1995, Carpenter et al. 1998, Tomasello et al. 2005).[15] This example illustrates concretely that investments in the formation of skills in one dimension of human capital can spill over into the formation of skills in a different dimension of human capital.

In sum, certain elements of human capital have two characteristics not shared by the types of skills, such as crystallized intelligence, that are captured by Equation 9. First, they are subject to critical or sensitive periods of development. Second, the investments in such skills formed in developmental stage $t$ spill over into (i.e., they complement or substitute the investments in) the development of skills produced in developmental stage $t + \tau$. For these elements of human capital, researchers turn to Cunha & Heckman's (2007) model of skill formation, whose specification is

$$h_{i,t+1} = g\left(x_{i,t}, x_{i,t-1}, \ldots, x_{i,1}, \eta_{i,t}\right). \qquad 10.$$

There are several noteworthy applications of these concepts. For example, researchers interested in the accumulation of human capital that takes place in utero investigate how these investments affect later health development (e.g., see the discussion of the evidence on this topic in Almond et al. 2018). Other researchers are interested in understanding if fostering secure attachment between parents and children influences later health and socioemotional development (as in the Nurse-Family Partnership program; Olds 2002). Developmental psychologists investigate if early investments in language development affect school performance on reading comprehension skills (e.g., Hart & Risley 1995, Gilkerson et al. 2018). Economists estimate the causal impact of executive functions developed in early childhood on adult outcomes (e.g., Heckman & Karapacula 2019a,b).

Under certain conditions described in their article, Cunha & Heckman (2007) show that it is possible to represent Equation 10 in a recursive fashion as

$$h_{i,t+1} = h\left(h_{i,t}, x_{i,t}, \eta_{i,t}\right). \qquad 11.$$

The recursive formulation eliminates a relevant aspect of Cunha & Heckman's (2007) approach. Significantly, under recursivity, the skill produced in the last period is a sufficient statistic for the development of the skill undergoing sensitive periods of development in the current period. This property may rule out long-lasting processes of human development.

The advantage of the recursive representation is that it is empirically tractable. It is rarely the case that researchers can observe the entire history of investments. Under recursivity, one can represent the technology of skill formation as a function only of the skill produced in the previous

[14] Joint attention is the ability to focus on the same thing (object, person, event) with another person. Therefore, there are three parties involved in joint attention: the child, the object of focus, and another person.

[15] Theory of mind is a social-cognitive skill that reflects the capacity to think about one's and others' mental states (desires, emotions, knowledge, and beliefs) and to recognize that others' mental states may be different from one's own.

period and current investments. Because of its tractability, the recursive representation dominates empirical work.

## 4.2. Three Challenges in the Estimation of Production Functions

In this subsection we discuss how the literature on the estimation of production functions of human capital has addressed the lack of cardinality of test scores, the errors in the measures of human capital and investments, and the correlation between observed and unobserved inputs.

**4.2.1. Anchoring.** The most pressing challenge in estimating human capital production functions relates to the fact that researchers need to work with measures of human capital and investments that are cardinal. A production function describes how the inputs ($h_{i,t}$ units of current-period human capital and $x_{i,t}$ units of current-period investments) combine to produce the output ($h_{i,t+1}$ units of next-period human capital). This combination involves adding or multiplying these inputs, operations that we can only perform if the variables are cardinal.

The typical solution to the lack of cardinality in standard skill measures in empirical applications is to anchor test scores to adult outcomes that do have a cardinal scale (e.g., Cunha et al. 2010). There is very little research to date testing the sensitivity of the resulting production function estimates to different anchoring outcomes. Cunha et al. (2010) use two different outcomes as anchors. One outcome relates to educational attainment, and the other relates to involvement with the criminal justice system. They report that the implied optimal early-to-late ratio of investments depends on the anchoring outcome because the two anchors have different relationships with the various skills in their analysis. Consistent with the arguments in Section 2, their findings suggest that the goal of the analysis should guide the choice of the anchoring outcome. For example, if researchers wish to assess the impact of different human capital investment strategies on labor income inequality, labor earnings should serve as anchors for this type of analysis. If the goal is to evaluate how adolescents' investments affect college enrollment, then the probability of enrolling should serve as the anchor.

**4.2.2. Measurement error.** Second, researchers need to address the fact that our measures of human capital and investments in early childhood suffer from measurement error. In economics, we measure investments in human capital by the value of goods and the opportunity cost of time that families use to increase a child's human capital. As a result, economists have tended to use measures of investments collected from a time-use survey (e.g., Del Boca et al. 2014), or composite scales that reflect a combination of goods and time, such as the HOME Inventory (e.g., Cunha et al. 2010), or both, either separately or combined (e.g., Attanasio et al. 2014, 2019). In contrast, developmental psychologists measure investments by the quality of parent-child interactions specific to the child's age and to the dimension of human capital under study (e.g., Roggman et al. 2013, Gilkerson et al. 2018). Measures of investments culled from time-use surveys and environment scales such as the HOME Inventory will have errors in capturing such forms of interaction.

Measuring human capital in early childhood is at least as challenging as measuring investments. There are two approaches to assessing child development. First, one can assess development directly, through the interaction between a trained assessor and the focus child. This type of assessment is the norm in many gold standard instruments, such as the Bayley Scales of Infant Development (Bayley 1969). Second, other instruments use parental reports [e.g., the Caregiver Reported Early Childhood Development (CREDI); McCoy et al. 2018]. Measurement error arises in both approaches because of the difficulty of assessing many elements of human capital given that children have limited language capacity [e.g., de Villiers & de Villiers (2014) discuss how children's language skills could affect the measurement of theory of mind skills]. Therefore, developmental instruments may leave out many critical dimensions of early childhood human capital simply because we do not know how to extract information from children whose communication skills are limited, not because these dimensions are not relevant. Additionally, parent-report scales may have notable reporting errors (Bennetts et al. 2016, Moens et al. 2018).

In estimating human capital production functions, researchers address measurement error by casting the estimation problem in terms of dynamic factor models (e.g., Cunha et al. 2010, Agostinelli & Wiswall 2016, Attanasio et al. 2020). Currently, there are four estimators of such models. Attanasio et al. (2019) use an adaptation of factor score regression to estimate the parameters of the technology of skill formation proposed by Heckman et al. (2010). Agostinelli & Wiswall (2016) explore the generalized method of moments (GMM). Cunha et al. (2010) use nonlinear filtering methods to address missing data for measures of human capital and investments in the Children and Young Adults cohort of the National Longitudinal Survey of Youth 1979 (NLSY79-C/YA). Attanasio et al. (2020) propose a three-step simulation algorithm. In the first step, they estimate moments of observed measures. In the second step, they match the moments of the observed measures to the moments imposed by the factor structure. In the third step, they simulate the factors and recover the production function parameters.

The four estimators described above share common assumptions—dedicated measurement systems, measurement errors, and endogenous investments—but differ in the assumptions they make on the error terms. The factor score and GMM approaches do not require assumptions about the distribution of the error terms. The filtering (both linear or nonlinear) and simulation approaches impose flexible parametric assumptions on the data (e.g., the density function is approximated by a mixture of normal density functions). Regardless of the estimation methodology, this research shows that not accounting for measurement errors has a nonnegligible impact on the production function estimates or the simulation of counterfactual policies.

#### 4.2.3. Correlation between observed inputs and unobserved terms.

Apart from cardinality and measurement error, researchers interested in estimating early childhood human capital production functions have to deal with the fact that observed inputs likely correlate with unobserved inputs. This challenge is similar to the one economists face in the estimation of production functions in other subdisciplines such as industrial organization (e.g., Olley & Pakes 1996). Current-period investments are likely to be endogenous, because they correlate with current-period unobserved error terms. This form of endogeneity is not the product of, but rather stands in addition to, measurement error. The endogeneity of investments would still exist even if we measured investments without error, and we would still have to address measurement error in investments even if we were able to allocate investments randomly across children.

The literature typically makes use of instrumental variables to derive consistent estimates of the parameters of human capital production functions. For example, Cunha et al. (2010) use innovations in income, Attanasio et al. (2020) explore variation in prices, and Attanasio et al. (2019) rely on variation in investments induced by random allocation to a parenting program. The validity of each of these instruments depends on the nature of the unobserved error terms. For example, suppose the unobserved terms capture omitted inputs. In that case, it is unlikely that any of these variables satisfies the condition to be an exclusion restriction, because inputs could react to changes in income, changes in prices, or assignment to the control or the treatment parenting program.

If the unobserved error terms capture omitted inputs that families can only change at significant cost (e.g., moving to a different neighborhood), then changes in the prices of investment goods or random assignment to the control group or the treatment parenting education group could satisfy the necessary exclusion restrictions. Whether income innovations would satisfy such conditions depends on whether they capture innovations in permanent or transitory income. If the innovations are in permanent income, families could adjust the levels of these semifixed inputs in response to permanent income changes, thereby invalidating the instrument.

Additionally, the use of income or prices as a source of exogenous variation may not be valid if the goods and services are low-quality measures of the investments in children's human capital. For example, if our goal is to measure investments in early language development, then we should attempt to measure the frequency of joint attention events. The number of such events in a given month may or may not correlate with the number of books the family purchases in that period. If not, then income or prices would not serve as an instrument for actual investment (number of joint attention events), even though they likely predict the number of books in the home.

Human capital production functions differ from other types of production functions because the lags of the dependent variable are inputs. Other forms of endogeneity will arise in these types of models if unobserved error terms exhibit serial correlation. The research to date has paid little attention to the fact that current-period human capital possibly correlates with unobserved error terms that have serial dependence. In theory, it is possible to design identification strategies that address such forms of endogeneity. In practice, existing data may not have all of the information that such methodologies require. Longitudinal studies such as the NLSY79-C/YA, the Child Development Supplement (CDS) of the Panel Study of Income Dynamics (PSID), or the Early Childhood Longitudinal Study Birth Cohort (ECLS-B) follow children and families for long periods. However, they may lack variables that satisfy the conditions for exclusion restriction. Data from experimental studies may have more viable candidates for instrumental variables but may follow children only for a few years. Therefore, new study protocols that produce high-quality exclusion restrictions in multiple periods would be necessary to shed light on the serial correlation of unobserved error terms in the production of human capital.

## 5. NEW RESEARCH DIRECTIONS

As our review makes clear, there has been significant progress in the development of the econometric tools economists use to understand human capital formation in early childhood. In this section, we identify some new directions for research in this field.

The majority of the empirical work in human development economics has recognized the importance of accounting for errors in human capital and investment measures. Therefore, the factor model has become this literature's workhorse, especially when the objective is to estimate a human capital production function. In this sense, there is active research on refining the identification conditions established by Cunha et al. (2010). For example, Agostinelli & Wiswall (2016) show how identification conditions can be relaxed when researchers are willing to impose additional restrictions on the shape of the production functions and have access to age-invariant measures of human capital. In a similar vein, Embrey (2019) relaxes the CES specification used by Cunha et al. (2010) and shows how more flexible representations of the production function significantly improve model fit.

The typical approach in the empirical literature is to assume that the measurement error is unbiased (i.e., that the errors have a mean of zero conditional on observable characteristics of the child). However, recent research has uncovered evidence that the errors may be biased, with the bias varying with the demographic characteristics of the study participants. For example, Segal (2012) showed that performance in a coding speed test varied according to incentive schemes. Notably, the individuals whose performance changed as a function of the incentive schemes tended to have low scores on a conscientiousness test. In early childhood, Charness et al. (2019) influenced a child's performance on a task that measured theory of mind by asking the child to think about an interaction with another child. Among children growing up in low-income households, this form of intervention substantially raised the performance of older children (i.e., enrolled in grade 1 or above) on the theory of mind task. We need to adapt these types of study protocols to other measures of child development. By doing so, we can build a body of evidence on the patterns of bias in measurement errors, thus providing direction for the development and application of new methods that can account for biased measurement error (e.g., Hu & Schennach 2008, Hoderlein & Winter 2010).

In economics, we commonly categorize human capital into broad constructs such as cognitive and noncognitive skills. This type of classification makes sense because different types of skills have diverse impacts on socioeconomic outcomes (see, e.g., Heckman et al. 2006). Our discussion of human capital formation offers, however, another distinction, between skills that individuals accumulate over time and are not subject to sensitive periods of development (with crystallized intelligence as one example) and skills that undergo sensitive periods of development that may occur in diverse stages of the life cycle (i.e., either in early childhood or adolescence). This alternative categorization matters for the specification of the human capital production function, and we need to develop research protocols so that our measures correspond to such constructs. For example, new research that links behavioral economics to human capital formation has started to investigate how strategic thinking evolves from early childhood to adulthood (e.g., Brocas & Carrilo 2018, 2020).

Concomitantly, we need to refine our measures of investments to be able to capture objective and detailed information about both the quality and the quantity of adult-child interactions, as research suggests that social interaction is essential for early human capital formation (e.g., Kuhl et al. 2003). For example, for early language development, joint attention and speech recasting are vital components of the interaction. Both of these components involve conversational turns between the child and an adult person. Cunha & Nihtianova (2020) show that the number of conversational turns a child experiences around age 1 year predicts significantly language development at age 2 years. Updating our investment measures matters not only for the estimation of the human capital production functions but also for the design of new interventions (e.g., Suskind et al. 2015, Cunha et al. 2020).

A better understanding of the processes driving human capital formation allows for new and better insights for public policy design. For example, there is no evidence of sensitive periods of development for crystallized skills. Therefore, early interventions that target these skills' formation will not necessarily lead to long-term impacts, because the control group could eventually catch up with the intervention group. However, if the intervention focuses on skills that undergo sensitive periods of development and targets children at risk of not developing such skills, then, theoretically, children in the control group may never catch up with the children in the intervention group.

As we incorporate new human capital measures into economic studies of early childhood development, new methods should ensure that these measures have cardinal scales. We envision a two-step approach. In the first step, one fixes the outcome of interest (e.g., later-life earnings). Different objectives/outcomes necessitate the construction of different scales. As discussed previously, cardinality is context dependent, and so must be the construction of cardinal scales. In the second step, one recreates the same process used in the current design of psychometric scales but with a different measurement objective. In some circumstances, one could take advantage of existing items and repurpose them to the new (economic) context, a method developed by Nielsen (2019). In this approach, one retains (or weighs more heavily) the items strongly associated with the outcome of interest and discards (or weighs less) those items that do not correlate with the outcome.

## 6. CONCLUSION

This review presented some vibrant developments in the econometrics of early childhood human capital formation. We showed how this research tackles the lack of cardinality in human capital measures and the ubiquity of measurement error in early childhood data, and we suggested guidance about how to choose between specifications of human capital production functions. Also, we showed how anchoring, measurement error, and the correlation between observed inputs and unobserved error terms affect estimates of the parameters of human capital production functions.

We suggested several directions for future research. In short, we called for refinements of the identification arguments for human capital production functions, and we summarized evidence pointing to the fact that measurement errors in such settings may not be classical.

We also presented new developments in the production of cardinal measures of human capital that closely link with the types of skills that undergo sensitive periods of development in early childhood. We further highlighted the need for new investment measures that better reflect the types of adult-child interactions that influence the development of such skills.

Improvements in methodology and in the measuring of human capital and investments will lead to a more complete understanding of skill development in early childhood. Better methods and data should yield new theoretical and empirical insights about which forms of constraints prevent families from investing in their children's human capital. Such progress will foster the design of new interventions, which, in turn, will inform new directions for public policy.

## DISCLOSURE STATEMENT

The analysis and conclusions set forth in this review are those of the authors and do not indicate concurrence by other members of the research staff, the Board of Governors, or the Federal Reserve System. The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

## ACKNOWLEDGMENTS

F.C. acknowledges support from the National Institutes of Health grant 1R01HD073221-01A1.

## LITERATURE CITED

Acemoglu D, Johnson S, Robinson JA. 2001. The colonial origins of comparative development: an empirical investigation. *Am. Econ. Rev.* 91(5):1369–401

Achenbach TM, Rescorla LA. 2001. *Manual for the ASEBA school-age forms & profiles*. Rep., Res. Cent. Child. Youth Fam., Univ. Vt., Burlington

Agostinelli F, Wiswall M. 2016. *Estimating the technology of children's skill formation*. NBER Work. Pap. 22442

Aigner DJ, Hsiao C, Kapteyn A, Wansbeek T. 1984. Latent variable models in econometrics. In *Handbook of Econometrics*, Vol. 2, ed. Z Griliches, MD Intriligator, pp. 1321–93. Amsterdam: Elsevier

Aizer A, Currie J, Simon P, Vivier P. 2018. Do low levels of blood lead reduce children's future test scores? *Am. Econ. J. Appl. Econ.* 10(1):307–41

Aizer A, Eli S, Ferrie J, Lleras-Muney A. 2016. The long-run impact of cash transfers to poor families. *Am. Econ. Rev.* 106(4):935–71

Almlund M, Duckworth AL, Heckman J, Kautz T. 2011. Personality psychology and economics. In *Handbook of the Economics of Education*, Vol. 4, ed. E Hanushek, S Machin, L Woessmann, pp. 1–181. Amsterdam: Elsevier

Almond D, Currie J, Duque V. 2018. Childhood circumstances and adult outcomes: act II. *J. Econ. Lit.* 56(4):1360–446

Anderson TW, Rubin H. 1956. Statistical inference in factor analysis. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 5, ed. J Neyman, pp. 111–50. Berkeley: Univ. Calif. Press

Aronson J. 2005. Biomarkers and surrogate endpoints. *Br. J. Clin. Pharmacol.* 59(5):491–94

Aslin RN. 2014. Phonetic category learning and its influence on speech production. *Ecol. Psychol.* 26(1–2):4–15

Attanasio O, Cattan S, Fitzsimons E, Meghir C, Rubio-Codina M. 2020. Estimating the production function for human capital: results from a randomized control trial in Colombia. *Am. Econ. Rev.* 110(1):48–85

Attanasio O, Cunha F, Jervis P. 2019. *Subjective parental beliefs: their measurement and role*. NBER Work. Pap. 26516

Attanasio O, Fernandes C, Fitzsimons EOA, Grantham-McGregor SM, Meghir C, Rubio-Codina M. 2014. Using the infrastructure of a conditional cash transfer program to deliver a scalable integrated early child development program in Colombia: cluster randomized controlled trial. *Br. Med. J.* 349:g5785

Barrett G, Donald S. 2003. Consistent tests for stochastic dominance. *Econometrica* 71:71–104

Bayley N. 1969. *Bayley Scales of Infant Development*. San Antonio, TX: Psychol. Corp.

Ben-Porath Y. 1967. The production of human capital and the life cycle of earnings. *J. Political Econ.* 75(4):352–65

Bennetts SK, Mensah FK, Westrupp EM, Hackworth NJ, Reilly S. 2016. The agreement between parent-reported and directly measured child language and parenting behaviors. *Front. Psychol.* 7:1710

Blanton H, Jaccard J. 2006. Arbitrary metrics in psychology. *Am. Psychol.* 62:27–41

Bond TN, Lang K. 2013. The evolution of the black-white test score gap in grades K-3: the fragility of results. *Rev. Econ. Stat.* 95:1468–79

Bond TN, Lang K. 2018. The black-white education scaled test-score gap in grades K-7. *J. Hum. Resourc.* 53(4):891–917

Bond TN, Lang K. 2019. The sad truth about happiness scales. *J. Political Econ.* 127(4):1629–40

Boneva T, Rauh C. 2018. Parental beliefs about returns to educational investments—the later the better? *J. Eur. Econ. Assoc.* 16(6):1669–711

Bonhomme S, Robin JM. 2009. Consistent noisy independent component analysis. *J. Econom.* 149(1):12–25

Brocas I, Carrilo J. 2018. The determinants of strategic thinking in preschool children. *PLOS ONE* 13(5). **https://doi.org/10.1371/journal.pone.0195456**

Brocas I, Carrilo J. 2020. The development of social strategic ignorance and other regarding behavior from childhood to adulthood. *J. Behav. Exp. Econ.* 85:101524

Carpenter M, Nagell K, Tomasello M. 1998. Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monogr. Soc. Res. Child Dev.* 63(4):1–143

Cattell RB. 1963. Theory of fluid and crystallized intelligence: a critical experiment. *J. Educ. Psychol.* 54(1):1–22

Caucutt EM, Lochner L. 2020. Early and late human capital investments, borrowing constraints, and the family. *J. Political Econ.* 128(3):1065–147

Cawley J, Heckman J, Vytlacil E. 1999. On policies to reward the value added by educators. *Rev. Econ. Stat.* 81(4):720–27

Charness G, List JA, Rustichini A, Samek A, Van De Ven J. 2019. Theory of mind among disadvantaged children: evidence from a field experiment. *J. Econ. Behav. Organ.* 166:174–94

Chen X, Hong H, Nekipelov D. 2011. Nonlinear models of measurement errors. *J. Econ. Lit.* 49(4):901–37

Chetty R, Friedman JN, Rockoff JE. 2014. Measuring the impacts of teachers II: teacher value-added and student outcomes in adulthood. *Am. Econ. Rev.* 104(9):2633–79

Chorniy A, Currie J, Sonchak L. 2020. Does prenatal WIC participation improve child outcomes? *Am. J. Health Econ.* 6(2):169–98

Chrisman NR. 1998. Rethinking levels of measurement for cartography. *Cartogr. Geogr. Inform. Syst.* 25(4):231–42

Croon M. 2002. Using predicted latent scores in general latent structure models. In *Latent Variables and Latent Structure Models*, ed. GA Marcoulides, I Moustaki, pp. 195–224. Mahwah, NJ: Lawrence Erlbaum

Cunha F, Elo I, Culhane J. 2013. *Eliciting maternal beliefs about the technology of skill formation*. NBER Work. Pap. 19144

Cunha F, Gerdes M, Nihtianova S. 2020. *Language environment and maternal expectations: an evaluation of the Lena Start program*. Work. Pap., Rice Univ., Houston, TX

Cunha F, Heckman JJ. 2007. The technology of skill formation. *Am. Econ. Rev.* 97(2):31–47

Cunha F, Heckman JJ. 2008. Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation. *J. Hum. Resourc.* 43(4):738–82

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

Cunha F, Nihtianova S. 2020. *Measuring early investments in language development*. Work. Pap., Rice Univ., Houston, TX

Cunha F, Wolpin KI. 2020. *An evaluation of the Alief Independent School District JumpStart program: using a model to recover mechanisms from an RCT*. Work. Pap., Tex. Policy Lab, Rice Univ., Houston

Currie J. 2013. Pollution and infant health. *Child Dev. Perspect.* 7(4):237–42

de Villiers JG, de Villiers PA. 2014. The role of language in theory of mind development. *Top. Lang. Disord.* 34(4):313–28

Del Boca D, Flinn C, Wiswall MJ. 2014. Household choices and child development. *Rev. Econ. Stud.* 81(1):137–85

Eimas PD, Miller JL. 1992. Organization in the perception of speech by young infants. *Psychol. Sci. Public Interest* 3(6):340–45

Embrey I. 2019. *Re-estimating the technology of cognitive and noncognitive skill formation*. Work. Pap., Lancaster Univ., Lancaster, UK

Gertler P, Heckman J, Pinto R, Zanolini A, Vermeersch C, et al. 2014. Labor market returns to an early childhood stimulation intervention in Jamaica. *Science* 344(6187):998–1001

Gilkerson J, Richards JA, Warren SF, Oller DK, Russo R, Vohr B. 2018. Language experience in the second year of life and language outcomes in late childhood. *Pediatrics* 142(4):e20174276

Hall RE, Jones CI. 1999. Why do some countries produce so much more output per worker than others? *Q. J. Econ.* 114(1):83–116

Hansen KT, Heckman JJ, Mullen KJ. 2004. The effect of schooling and ability on achievement test scores. *J. Econom.* 121(1–2):39–98

Hart B, Risley T. 1995. *Meaningful Differences in the Everyday Experience of Young American Children*. Baltimore, MD: Paul H. Brookes

Heckman JJ, Karapacula G. 2019a. *Intergenerational and intragenerational externalities of the Perry Preschool Project*. NBER Work. Pap. 25889

Heckman JJ, Karapacula G. 2019b. *The Perry preschoolers at late midlife: a study in design-specific inference*. HCEO Work. Pap. 2019-034, Univ. Chicago, Chicago

Heckman JJ, Lochner L, Taber C. 1998. Explaining rising wage inequality: explorations with a dynamic general equilibrium model of labor earnings with heterogeneous agents. *Rev. Econ. Dyn.* 1(1):1–58

Heckman JJ, Mosso S. 2014. The economics of human development and social mobility. *Annu. Rev. Econ.* 6:689–733

Heckman JJ, Pinto R, Savelyev P. 2013. Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *Am. Econ. Rev.* 103(6):2052–86

Heckman JJ, Schennach MS, Williams B. 2010. *Matching with error-laden covariates*. Unpublished manuscript, Univ. Chicago, Chicago

Heckman JJ, Stixrud J, Urzua S. 2006. The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *J. Lab. Econ.* 24(3):411–82

Ho A. 2009. A nonparametric framework for comparing trends and gaps across tests. *J. Educ. Behav. Stat.* 34:201–28

Ho A, Quinn D. 2020. *Ordinal approaches to decomposing between-group test score disparities*. Work. Pap. 20-257, Annenberg Inst., Brown Univ., Providence, RI

Ho A, Reardon S. 2012. Estimating achievement gaps from test scores reported in ordinal "proficiency" categories. *J. Educ. Behav. Stat.* 37(4):489–517

Hoderlein S, Winter J. 2010. Structural measurement errors in nonseparable models. *J. Econom.* 157:432–40

Hu Y, Schennach SM. 2008. Instrumental variable treatment of nonclassical measurement error models. *Econometrica* 76(2):195–216

Jacob B, Rothstein J. 2016. The measurement of student ability in modern assessment systems. *J. Econ. Perspect.* 30:85–108

Joreskog KG, Goldberger AS. 1975. Estimation of a model with multiple indicators and multiple causes of a single latent variable. *J. Am. Stat. Assoc.* 70:631–39

Junker B, Schofield LS, Taylor LJ. 2012. The use of cognitive ability measures as explanatory variables in regression analysis. *IZA J. Lab. Econ.* 1(1):4

Kalil A. 2014. Inequality begins at home: the role of parenting in the diverging destinies of rich and poor children. In *Families in an Era of Increasing Inequality: Diverging Destinies*, ed. PR Amato, A Booth, SM McHale, J Van Hook, pp. 63–82. New York: Springer

Kuhl PK. 2004. Early language acquisition: cracking the speech code. *Nat. Rev. Neurosci.* 5:831–43

Kuhl PK, Tsao FM, Liu HM. 2003. Foreign-language experience in infancy: effects of short-term exposure and social interaction on phonetic learning. *PNAS* 100(15):9096–101

Kuhl PK, Williams KA, Lacerda F, Stevens KN, Lindblom B. 1992. Linguistic experience alters phonetic perception in infants by 6 months of age. *Science* 255(5044):606

La Porta R, Lopez-de Silanes F, Shleifer A, Vishny R. 1999. The quality of government. *J. Law Econ. Organ.* 15(1):222–79

Lang K. 2010. Measurement matters: perspectives on education policy from an economist and school board member. *J. Econ. Perspect.* 24:167–81

Lockwood JR, McCaffrey D. 2014. Correcting for test score measurement error in ANCOVA models for estimating treatment effects. *J. Educ. Behav. Stat.* 39(1):22–52

Lord F. 1975. The "ability" scale in item characteristics curve theory. *Psychometrika* 40:205–17

Lucas RE. 1988. On the mechanics of economic development. *J. Monet. Econ.* 22(1):3–42

Mauro P. 1995. Corruption and growth. *Q. J. Econ.* 110(3):681–712

McCoy DC, Waldman M, Altafim E, Brentani A, Castellanos A, et al. 2018. Measuring early childhood development at a global scale: evidence from the caregiver-reported early development instruments. *Early Childh. Res. Q.* 45:58–68

Moens MA, Weeland J, Giessen DVd, Chhangur RR, Overbeek G. 2018. In the eye of the beholder? Parent-observer discrepancies in parenting and child disruptive behavior assessments. *J. Abnorm. Child Psychol.* 46:1147–59

Mosteller F, Turkey JW. 1977. *Data Analysis and Regression: A Second Course in Statistics*. Reading, MA: Addison-Wesley

Nielsen E. 2015a. *Achievement gap estimates and deviations from cardinal comparability*. Finance Econ. Discuss. Ser. 2015-040, Fed. Reserve Board, Washington, DC

Nielsen E. 2015b. *The income-achievement gap and adult outcome inequality*. Finance Econ. Discuss. Ser. 2015-041, Fed. Reserve Board, Washington, DC

Nielsen E. 2019. *Test questions, economic outcomes, and inequality*. Finance Econ. Discuss. Ser. 2019-013, Fed. Reserve Board, Washington, DC

Olds D. 2002. Prenatal and infancy home visiting by nurses: from randomized trials to community replication. *Prev. Sci.* 3(3):153–72

Olley S, Pakes A. 1996. The dynamics of productivity in the telecommunications equipment industry. *Econometrica* 64(6):1263–97

Polachek SW, Das T, Thamma-Apiroam R. 2016. Micro- and macroeconomic implications of heterogeneity in the production of human capital. *J. Political Econ.* 123(6):1410–55

Reiersol O. 1950. On the identifiability of parameters in Thurstone's multiple factor analysis. *Psychometrika* 15(2):121–49

Roggman LA, Cook GA, Innocenti MS, Norman VJ, Christiansen K. 2013. Parenting interactions with children: checklist of observations linked to outcomes (PICCOLO) in diverse ethnic groups. *Infant Mental Health J.* 34(4):290–306

Rothstein J. 2010. Teacher quality in educational production: tracking, decay, and student achievement. *Q. J. Econ.* 125(1):175–214

Schennach SM. 2016. Recent advances in the measurement error literature. *Annu. Rev. Econ.* 8:341–77

Schofield L. 2014. Measurement error in the AFQT in the NLSY79. *Econ. Lett.* 123(3):262–65

Schröder C, Yitzhaki S. 2017. Revisiting the evidence for cardinal treatment of ordinal variables. *Eur. Econ. Rev.* 92:337–58

Segal C. 2012. Working when no one is watching: motivation, test scores, and economic success. *Manag. Sci.* 58:1438–57

Sijtsma K, Junker BW. 2006. Item response theory: past performance, present developments, and future expectations. *Behaviormetrika* 33(1):75–102

Stevens S. 1946. On the theory of scales of measurement. *Science* 103:677–80

Suskind DL, Leffel KR, Graf E, Hernandez MW, Gunderson EA, et al. 2015. A parent-directed language intervention for children of low socioeconomic status: a randomized controlled pilot study. *J. Child Lang.* 43(2):366–406

Thorndike RL. 1951. Reliability. In *Educational Measurement*, ed. EF Lindquist, RL Thorndike, pp. 560–620. Washington, DC: Am. Counc. Educ.

Todd PE, Wolpin KI. 2003. On the specification and estimation of the production function for cognitive achievement. *Econ. J.* 113(485):F3–33

Todd PE, Wolpin KI. 2007. Production of cognitive achievement in children: home, school, and racial test score gaps. *J. Hum. Capital* 1(1):91–113

Tomasello M. 1995. Joint attention as social cognition. In *Joint Attention: Its Origins and Role in Development*, ed. C Moore, PJ Dunham, pp. 103–30. Mahwah, NJ: Lawrence Erlbaum

Tomasello M, Carpenter M, Call J, Behne T, Moll H. 2005. Understanding and sharing intentions: the origins of cultural cognition. *Behav. Brain Sci.* 28(5):675–91

van der Linden WJ, Hambleton RK. 2013. *Handbook of Modern Item Response Theory*. New York: Springer

Wansbeek T, Meijer E. 2001. *Measurement Error and Latent Variables*. Amsterdam: North Holland

Werker JF, Gilbert JHV, Humphrey K, Tees RC. 1981. Developmental aspects of cross-language speech perception. *Child Dev. Perspect.* 52(1):349–55

Werker JF, Tees RC. 2005. Speech perception as a window for understanding plasticity and commitment in language systems of the brain. *Dev. Psychobiol.* 46:233–51

Williams B. 2019. Controlling for ability using test scores. *J. Appl. Econom.* 34(4):547–65

Williams B. 2020. Identification of the linear factor model. *Econom. Rev.* 39(1):92–109

Annual Review of Economics

Volume 13, 2021

# Contents

## Indexes

## Errata

An online log of corrections to *Annual Review of Economics* articles may be found at http://www.annualreviews.org/errata/economics