

Normalizations and misspecification in skill formation models*

Joachim Freyberger[‡]

August 2025[†]

Abstract

An important class of structural models studies the determinants of skill formation and the optimal timing of interventions. In this paper, I provide new identification results for these models and investigate the effects of seemingly innocuous scale and location restrictions on parameters of interest. To do so, I first characterize the identified set of all parameters without these additional restrictions and show that important policy-relevant parameters are point identified under weaker assumptions than commonly used in the literature. The implications of imposing standard scale and location restrictions depend on how the model is specified, but they generally impact the interpretation of parameters and may affect counterfactuals. Importantly, with the popular CES production function, commonly used scale restrictions fix identified parameters and lead to misspecification. Consequently, simply changing the units of measurements of observed variables might yield ineffective investment strategies and misleading policy recommendations. I show how existing estimators can easily be adapted to solve these issues. As a byproduct, this paper also presents a general and formal definition of when restrictions are truly normalizations.

*Thanks to Antonia Antweiler, Gabriella Conti, Philipp Eisenhauer, Janos Gabler, Hans-Martin von Gaudecker, Lena Janys, Julius Kappenberg, Philipp Ketz, Josh Kinsler, Matt Masten, Mauricio Olivares, Ronni Pavan, Jack Porter and seminar participants at TSE, UCL, PSE, LMU, Zurich, Luxembourg, Tsinghua, UCSD, McMaster, Yale, and the German Economic Association for helpful comments and to Matt Wiswall for many useful discussions. I also thank five referees and the editor for constructive comments that helped to improve the paper. The research was supported by the European Research Council (ERC-2020-STG-949319).

[‡]University of Bonn. Email: freyberger@uni-bonn.de.

[†]First version: April 10, 2020

1 Introduction

Structural models are key tools of economists to simulate changes in the economic environment and evaluate and design policies. An important class of such models deals with skill and human capital formation. Human capital formation is a main part of many structural models and is an important driver of economic growth and inequality (Murphy and Topel 2016), making policies that target skill formation particularly vital. This growing literature, originating from the seminal papers of Cunha and Heckman (2008) and Cunha, Heckman, and Schennach (2010), estimates production functions of various skills of children, studies how past skills, parental skills, and investments affect future skills, and links the skills to adult outcomes. The results provide valuable insights into determinants of skill formation, timing of investments, and the design of optimal interventions for disadvantaged children.

A major challenge in these models is that skills are not directly observable, lack a natural scale, and can only be approximated through measurements, such as test scores. The early literature has provided sufficient conditions for identification of parametric and nonparametric versions of these complex models, which is often achieved using a two-step approach. First, the distribution of skills is identified from the measurements, and the production function is then identified in the second step (see e.g. Cunha et al. (2010), Attanasio, Meghir, Nix, and Salvati (2017), and Attanasio, Meghir, and Nix (2020)). To obtain point identification in the first step, it is necessary to fix the unknown scales and locations of skills. However, it remains unclear whether these restrictions are still required when combined with common parametric assumptions on the production function in the second step.

This paper presents a new identification analysis for skill formation models and investigate the consequences of seemingly innocuous normalizations. Instead of providing sufficient conditions for point identification using multi-step arguments, I start by pooling all parts of the model and characterize the identified set of all parameters without the scale and location restrictions. This approach reveals which parameters are point identified or partially identified and how the additional restrictions affect the identified set. Notably, I show that many critical features of these models are invariant to scale and location restrictions, point-identified without them, and identified under less restrictive assumptions than previously

considered. These results apply to both parametric and nonparametric versions of the model.

The exact implications of imposing scale and location restrictions on parameters and counterfactuals depend on the model specification and the object of interest. Specifically, a restriction could be a harmless normalization with one production function, but could impose strong assumptions with a different one. I thus analyze two popular specifications, namely the trans-log and the CES production functions. For the trans-log case, it turns out that standard scale and location restrictions do not impose additional testable restrictions, simply select an element of the identified set, and many objects of interest are invariant to them. However, they still affect production function parameters and certain counterfactuals. Since such restrictions are often arbitrary and tied to the units of measurement of the data, these features can be hard to interpret, difficult to compare across different studies, and potentially mislead policy recommendations. More importantly, for the CES case, I show that commonly restricted scale parameters are in fact identified and setting them to specific values is typically inconsistent with the data. Consequently, with these restrictions, simply changing the units of a skill measure (e.g. from years to months) can impact estimated dynamics, persistence of skills, effects of parental investments, and optimal investment strategies.

In addition to these identification results, I demonstrate how existing estimators can be modified to estimate all identified parameters. While some parameters remain difficult to interpret, the modified estimator enables the calculation of point identified features and counterfactuals that are invariant to scale and location restrictions and to the units of measurement. I illustrate these results through Monte Carlo simulations and an empirical application based on the framework and estimator of Attanasio et al. (2020).

A broader takeaway is that researchers should carefully verify whether imposed restrictions are in fact normalizations (as formally defined in Definition 1). This is particularly crucial in structural models where identification often proceeds in multiple steps, and supposedly innocuous restrictions in one step may unintentionally affect results in subsequent steps. By focusing on features of the model that are invariant to these restrictions, researchers can ensure that their findings remain interpretable and comparable across different studies.

Literature: Following the influential work of Cunha and Heckman (2008) and Cunha et al. (2010), a growing body of literature has studied the development of latent variables.

For example, Helmers and Patnam (2011) study the determinants of children’s cognitive and non-cognitive skills using Indian data, Fiorini and Keane (2014) investigate how time allocation affects both cognitive and non-cognitive development using Australian data, Attanasio et al. (2020) and Attanasio et al. (2017) estimate the effects of health and cognition on human capital with data from India and Ethiopia/Peru, respectively, and Attanasio et al. (2020) study which intervention led to gains in cognitive and socio-emotional skills using Colombian data. This literature provides important insights into the nature of persistence, dynamic complementarities, and the optimal targeting of interventions.¹

In these studies, the measurement system has a factor structure that requires scale and location restrictions to identify the distribution of latent variables. In particular, Cunha and Heckman (2008) study a model where the production function is log-linear. They utilize the two-step identification arguments described above, except that they impose the scale and location restrictions using an adult outcome to obtain well-defined units of measurement (“anchoring”). They also discuss which parameters depend on the anchor or its units of measurement. In this model, I demonstrate that certain policy-relevant parameters, such as optimal investment sequences, may depend on the specific scales and units of measurements of the observed variables. I further derive a set of features that is identified under weaker restrictions on the measures and the production function and without anchoring the skills or fixing their scales. Importantly, I also provide identification results for alternative production technologies, such as the widely used CES production function, where standard restrictions are unnecessary for identification. In these case, imposing them through an initial period normalization or through anchoring leads to misspecification.

Common restrictions are to fix the scales and locations of each latent factor in each time period by setting parameters in the measurement system.² If the same values are

¹See also Cunha and Heckman (2007), Cunha and Heckman (2009), Cunha (2011), Heckman, Pinto, and Savelyev (2013), Aucejo and James (2021), Hernández-Alava and Popli (2017), Attanasio, Cattani, and Meghir (2022) and references therein.

²For example, Cunha and Heckman (2008) and Cunha et al. (2010) set the scale of the first skill measure to 1 in each period - see the discussion around equations (7) – (8) of Cunha and Heckman (2008) and the first paragraph on page 891 of Cunha et al. (2010). In addition, as discussed in footnote 17 of Cunha and Heckman (2008), their identification results rely on either setting one of the location parameters to 0 in each time period or setting the mean of the skill to 0 in each time period. These restrictions have also been used in subsequent papers such as Attanasio et al. (2020). The nonparametric identification results in Cunha et al. (2010) impose analogous restrictions in their Assumption (v) of Theorem 2.

used across all time periods, Agostinelli and Wiswall (2016a, 2016b, 2024) refer to this as an age-invariance assumption (see Assumption 3(a) for a formal definition or Definition 1 of Agostinelli and Wiswall (2024)). In important contributions, Agostinelli and Wiswall (2016a, 2016b, 2024) demonstrate that imposing this assumption can yield a misspecified model when the production function has a known scale and location. They also propose relaxations of production functions with age-invariant measures and show that, even without age-invariance, production function restrictions can yield point identification (see Corollary 1 below and the related discussion). However, in both cases, they impose scale and location restrictions in the first period, arguing that these are necessary for point identification. While this is true for the trans-log production function used in their application, I show that the scale restriction is not required for the CES production function. Furthermore, I demonstrate for different specifications that while production function parameters, certain counterfactuals, and estimated dynamics may depend on specific scales and units of measurements, many key parameters are invariant to these restrictions, including age-invariance, and are in fact point identified without them.

In independent research, Del Bono, Kinsler, and Pavan (2022) show that with a trans-log production function, anchored treatment effects are invariant to scale and location restrictions and are identified without age-invariance. Through simulations, they also show that standard restrictions lead to inconsistent estimated treatment effects with the CES production function. In the trans-log case, these results aligns with part 4 of Theorem 2 below. While their proof is specific to the trans-log production function with a log-linear measurement system, my results extend to other cases. Additionally, I show that only skill measures in the first period are needed to identify anchored treatment effects, reducing the data requirements and assumptions considerably. I also consider other policy-relevant features. Additionally, for the CES case, I show why standard restrictions can cause misspecification and describe how existing estimators can be adapted to resolve this issue.

Table 1 of Attanasio et al. (2022) provides a selective overview of specifications and estimation methods used in well-cited papers in the literature. It shows that with the exception of Cunha et al. (2010), earlier work predominantly employed the Cobb-Douglas specification, likely due to the simplicity of the corresponding estimator. Since Attanasio

et al. (2020), there has been increasing adoption of the (nested) CES production function, facilitated by their computationally simple two-step estimator (compared to the MLE of Cunha et al. (2010)). Next to the papers in this table, recent contributions using this approach include Adhvaryu et al. (2023), Bolt et al. (2024), and Gallipoli and Gomez (2023). The trans-log production function proposed by Agostinelli and Wiswall (2024) is nonnested with the CES specification and can also be estimated using an IV/GMM approach.

The consequences of normalizations have been discussed in various contexts. In factor models, while certain restrictions are necessary for point identification (see e.g. Anderson and Rubin (1956) or Madansky (1964)), Williams (2020) shows that certain features, such as variance decompositions, can be identified without them. I combine a factor model with a production function, which provides additional restrictions. Many studies argue and show in specific examples that critical features should not depend on normalizations, see e.g. Freyberger (2018) and Komarova et al. (2018). Similar to this paper, but in a very different context, Aguirregabiria and Suzuki (2014) discuss restrictions that were considered normalizations, but are restrictive assumptions. Kalouptsi et al. (2021) show that certain counterfactuals in dynamic discrete choice models are identified, even when the model itself is not. Rubio-Ramírez et al. (2010) define a normalization in vector autoregressive models to pin down unidentified signs; see end of Section 2 for more details. Matzkin (1994, 2007) discusses several examples of normalizations, some of which are motivated by economic theory. Lewbel (2019) provides an informal discussion of normalizations, which is conceptually very similar to the formal definition I provide below. When normalizing restrictions are needed for point identification, there are often multiple ways to impose them when estimating the model. Good choices can then yield particularly convenient restrictions on the parameter space (as in Gao and Li (2021)) or even faster rates of convergence (as in Chiappori et al. (2015)). See also Hamilton et al. (2007) for a discussion on estimation with normalizations.

Structure: In Section 2, I provide a formal definition of a normalization, which to the best my knowledge currently does not exist in the literature, as well as illustrative examples. Section 3 contains the identification analysis of different parametric skill formation models. A nonparametric version is discussed in Appendix D. Sections 4 and 5 contain the Monte Carlo simulations and the empirical application, respectively. All proofs are in the appendix.

2 Normalizations

I begin by providing a formal definition of a normalization, which serves as a basis for the subsequent analysis. I illustrate the definition and potential problems using a probit model and a simple version of the skill formation model.

2.1 General definition and illustration

Suppose we have a model where $\tau_0 \in \mathcal{T}$ denotes the true values of the parameters and \mathcal{T} is the parameter space. Here τ_0 could be the coefficients in a regression model or the parameters in a skill formation model. If the model is semiparametric, τ_0 could also contain unknown functions. Let Z contain all observed random variables, such as Y and X , with distribution $P(Z)$. For any $\tau \in \mathcal{T}$, the model generates a joint distribution of the data Z , denoted by $P(Z, \tau)$. Since the model is assumed to be correctly specified, the true distribution of Z is $P(Z, \tau_0)$. The model typically contains certain assumptions, such as functional form or independence assumptions, but suppose that so far none of the normalizations are imposed. The identified set for τ_0 is $\mathcal{T}_0 = \{\tau \in \mathcal{T} : P(Z, \tau) = P(Z, \tau_0)\}$. If \mathcal{T}_0 is a singleton, τ_0 is point identified. We say that $\tau_1, \tau_2 \in \mathcal{T}$ are observationally equivalent if they generate the same distribution of the data: $P(Z, \tau_1) = P(Z, \tau_2)$. Let $g(\tau_0)$ be a function of interest, such as a counterfactual. The identified set for $g(\tau_0)$ is $\mathcal{T}_{g_0} = \{g(\tau) : \tau \in \mathcal{T}_0\}$. Notice that $g(\tau_0)$ could be point identified (i.e. \mathcal{T}_{g_0} is a singleton) even if τ_0 is not.

In models with normalizations, \mathcal{T}_0 is typically not a singleton. A normalization is a restriction of the form $\tau \in \mathcal{T}_N$, where $\mathcal{T}_N \subseteq \mathcal{T}$ is a known set. Hence, a normalization restricts the feasible values of τ , such as setting an element to 1. I define a restriction to be a normalization with respect to a function $g(\tau_0)$ if it does not change the identified set of $g(\tau_0)$.

Definition 1. The restriction $\tau \in \mathcal{T}_N$ is a *normalization* with respect to $g(\tau_0)$ if $\{g(\tau) : \tau \in \mathcal{T}_0 \cap \mathcal{T}_N\} = \{g(\tau) : \tau \in \mathcal{T}_0\}$ for all $\tau_0 \in \mathcal{T}$.

Typically, $\mathcal{T}_0 \cap \mathcal{T}_N$ is a singleton. That is, we achieve point identification with the additional restrictions. Definition 1 then implies $\{g(\tau_0)\} = \{g(\tau) : \tau \in \mathcal{T}_0 \cap \mathcal{T}_N\} = \{g(\tau) : \tau \in \mathcal{T}_0\}$

and thus, that $g(\tau_0)$ is point identified, even without the restriction $\tau \in \mathcal{T}_N$. Since these restrictions are often arbitrary, τ_0 is usually not in \mathcal{T}_N in which case the restriction $\tau \in \mathcal{T}_N$ is not a normalization with respect to τ_0 , but it can be a normalization with respect to particular functions of interest. Moreover, the restriction can be a normalization for some function and not for others. Hence, researchers need to argue that normalizations hold with respect to all functions of interest, such as all counterfactuals. Finally, a normalization cannot impose any additional overidentifying restrictions in the sense that if $\mathcal{T}_0 \neq \emptyset$, then $\mathcal{T}_0 \cap \mathcal{T}_N \neq \emptyset$.

As a simple example, consider the probit model where $Y = \mathbf{1}(\beta_{0,1} + \beta_{0,2}X \geq U)$, $\text{var}(X) > 0$, $U | X \sim N(\mu_0, \sigma_0^2)$ and $\sigma_0^2 > 0$. The true parameter vector is $\tau_0 = (\beta_{0,1}, \beta_{0,2}, \mu_0, \sigma_0)'$ and $Z = (Y, X)$. Now notice that

$$P(Y = 1 | X = x) = \Phi \left(\frac{\beta_{0,1} - \mu_0}{\sigma_0} + \frac{\beta_{0,2}}{\sigma_0} x \right),$$

where Φ denotes the standard normal cdf. Since $\text{var}(X) > 0$, $\frac{\beta_{0,1} - \mu_0}{\sigma_0}$ and $\frac{\beta_{0,2}}{\sigma_0}$ are point identified. It is also well known and easy to see that

$$\mathcal{T}_0 = \left\{ \tau \in \mathbb{R}^3 \times \mathbb{R}_{>0} : \frac{\beta_1 - \mu}{\sigma} = \frac{\beta_{0,1} - \mu_0}{\sigma_0} \text{ and } \frac{\beta_2}{\sigma} = \frac{\beta_{0,2}}{\sigma_0} \right\}$$

because all values in \mathcal{T}_0 imply the same joint distribution of (Y, X) .

Since τ_0 is not point identified, it is common to set $\mu = 0$ and $\sigma = 1$. Using the previous notation, this means that $\mathcal{T}_N = \mathbb{R}^2 \times 0 \times 1$ and $\mathcal{T}_0 \cap \mathcal{T}_N = \left(\frac{\beta_{0,1} - \mu_0}{\sigma_0}, \frac{\beta_{0,2}}{\sigma_0}, 0, 1 \right)$. Clearly, this restriction is not a normalization with respect to $\beta_{0,1}$ or $\beta_{0,2}$, which are typically not objects of interest. In fact, in general $\tau_0 \notin \mathcal{T}_0 \cap \mathcal{T}_N$ unless $\mu_0 = 0$ and $\sigma_0 = 1$. However, this restriction is a normalization with respect to (potentially counterfactual) probabilities

$$P(Y = 1 | X = x) = \Phi \left(\frac{\beta_{0,1} - \mu_0}{\sigma_0} + \frac{\beta_{0,2}}{\sigma_0} x \right)$$

or, when X is continuous, marginal effects $\frac{\partial}{\partial x} P(Y = 1 | X = x)$. Clearly, these features are point identified even though τ_0 is not.

Normalizations may help to provide useful structural interpretations of other parameters. As an example, suppose there are two covariates, $Y = \mathbf{1}(\beta_{0,1} + \beta_{0,2}X_1 + \beta_{0,3}X_2 \geq U)$, and

$\beta_{0,2} > 0$. Instead of setting $\mu = 0$ and $\sigma = 1$, we could impose $\beta_{0,1} = 0$ and $\beta_{0,2} = 1$, which are also normalizations with respect to marginal effects. In addition, we then obtain a model of the form $Y = \mathbf{1}(X_1 + \tilde{\beta}_{0,3}X_2 \geq \tilde{U})$ with $\tilde{\beta}_{0,3} = \beta_{0,3}/\beta_{0,2}$, which can then be interpreted as a relative effect (which is also identified without scale and location restrictions).

In the context of vector autoregressive models, Rubio-Ramírez et al. (2010) define a normalization as a restriction on the parameter space that pins down unidentified signs of coefficients. These restrictions are imposed in addition to other assumptions, such as long run restrictions. Just like above, their restrictions do not impose additional testable assumptions and they can help to provide structural interpretations of certain parameters. Unlike my definition, it is not clear from theirs whether these restrictions are without loss of generality in the sense that they do not affect functions of interest. If they do, one would have to argue why they are reasonable.

2.2 Simple skill formation model

As a more involved example, I now discuss a very simple skill formation model, which imposes very restrictive assumptions, but illustrates the previous definition and points out the types of problems that occur in more general models.

Let θ_t denote skills at time t and let I_t be investment at time t , where $t = 0, 1, \dots, T$. We are interested in the roles of investment and past skills in the development of future skills. Instead of skills, the data only contains measurements of them, denoted by $Z_{\theta,t,m}$. In this section, I first consider the simplest possible model without measurement error and a single measure $Z_{\theta,t,1}$ that takes the form $Z_{\theta,t,1} = \lambda_{\theta,t,1} \ln \theta_t$, where $\{\lambda_{\theta,t,1}\}_{t=0}^T$ are unknown parameters with $\lambda_{\theta,t,1} \neq 0$ for all t . In other words, in each period, we observe a scaled version of log-skills. I also assume that there are three periods ($T = 2$) and that investment is observed. Finally, for simplicity, I parameterize the marginal distribution of skills in the initial period as $\ln \theta_0 \sim N(0, s_0^2)$, which implies that $E[Z_{\theta,0,1}] = 0$ and $Var(Z_{\theta,0,1}) = \lambda_{\theta,0,1}^2 s_0^2$.

I consider two simple production functions without unobserved random variables. First suppose skills evolve based on the Cobb-Douglas production function:

$$\ln \theta_{t+1} = a_t + \gamma_{1t} \ln \theta_t + \gamma_{2t} \ln I_t.$$

Define the vector containing the true values of all ten parameters as

$$\tau_0 = (\tau_{0,1}, \tau_{0,2}, \dots, \tau_{0,10}) = (\lambda_{\theta,0,1}, \lambda_{\theta,1,1}, \lambda_{\theta,2,1}, a_0, a_1, \gamma_{10}, \gamma_{20}, \gamma_{11}, \gamma_{21}, s_0^2).$$

Without further assumptions, τ_0 is not point identified. To see the restrictions imposed by the observables, notice that $\ln \theta_t = Z_{\theta,t,1}/\lambda_{\theta,t,1}$ and therefore

$$Z_{\theta,t+1,1} = \lambda_{\theta,t+1,1}a_t + \frac{\lambda_{\theta,t+1,1}}{\lambda_{\theta,t,1}}\gamma_{1t}Z_{\theta,t,1} + \lambda_{\theta,t+1,1}\gamma_{2t}\ln I_t.$$

Hence, the joint distribution of $\{Z_{\theta,t+1,1}, Z_{\theta,t,1}, I_t\}_{t=0}^1$ identifies (a) the scaled intercepts of the production function $\lambda_{\theta,t+1,1}a_t$, (b) the scaled slope coefficients $\frac{\lambda_{\theta,t+1,1}}{\lambda_{\theta,t,1}}\gamma_{1t}$ and $\lambda_{\theta,t+1,1}\gamma_{2t}$, and (c) the scale variance $\lambda_{\theta,0,1}^2 s_0^2$. The identified set can be shown to consist of all parameters that imply the same values of these identified features as the true parameters. Formally,

$$\begin{aligned} \mathcal{T}_0 = \left\{ \tau \in \mathcal{T} : \tau_2\tau_4 = \tau_{0,2}\tau_{0,4}, \tau_3\tau_5 = \tau_{0,3}\tau_{0,5}, \right. \\ \left. \frac{\tau_2}{\tau_1}\tau_6 = \frac{\tau_{0,2}}{\tau_{0,1}}\tau_{0,6}, \frac{\tau_3}{\tau_2}\tau_8 = \frac{\tau_{0,3}}{\tau_{0,2}}\tau_{0,8}, \tau_2\tau_7 = \tau_{0,2}\tau_{0,7}, \tau_3\tau_9 = \tau_{0,3}\tau_{0,9}, \right. \\ \left. \tau_1^2\tau_{10} = \tau_{0,1}^2\tau_{0,10} \right\} \end{aligned}$$

where the three rows correspond to the three sets of features above. To achieve point identification, we could set $\lambda_{\theta,t,1} = 1$ for all t , in which case the intersection of \mathcal{T}_0 and the additional restrictions is the singleton $\{(1, 1, 1, \tau_{0,2}\tau_{0,4}, \tau_{0,3}\tau_{0,5}, \frac{\tau_{0,2}}{\tau_{0,1}}\tau_{0,6}, \tau_{0,2}\tau_{0,7}, \frac{\tau_{0,3}}{\tau_{0,2}}\tau_{0,8}, \tau_{0,3}\tau_{0,9}, \tau_{0,1}^2\tau_{0,10})\}$.

As a simple numerical example, suppose $Z_{\theta,t,1} = 12 \ln \theta_t$ and

$$\ln \theta_{t+1} = 0.5 \ln \theta_t + 0.5 \ln I_t$$

for all t , and $\ln \theta_0 \sim N(0, 1)$. Here $\tau_0 = (12, 12, 12, 0, 0, 0.5, 0.5, 0.5, 0.5, 1)$. If we set $\lambda_{\theta,t,1} = 1$, even though the true value is 12, we essentially treat $\ln \tilde{\theta}_t \equiv Z_{\theta,t,1} = 12 \ln \theta_t$ as the skills and the corresponding production function is

$$\ln \tilde{\theta}_{t+1} = 0.5 \ln \tilde{\theta}_t + 0.5 \ln I_t.$$

The intersections of the identified set and set of additional restrictions is then the singleton $\{(1, 1, 1, 0, 0, 0.5, 6, 0.5, 6, 144)\}$ which would be the parameter estimated in practice (instead of τ_0). The coefficients in front of investment are thus hard to interpret. For example, using τ_0 (or setting $\lambda_{\theta,t,1} = 12$) one might conclude that increasing investment by 1% increases skills by 0.5%, but with the restriction $\lambda_{\theta,t,1} = 1$ that effect changes to 6%. Hence, $\lambda_{\theta,t,1} = 1$ is not a normalization with respect to these parameters. The coefficient in front of log-skills is invariant to the scale restrictions, but only because here $\lambda_{\theta,t,1} = \lambda_{\theta,t+1,1}$ for all t . Notice that $Z_{\theta,t,1}$ is simply a scaled version of $\ln \theta_t$. If we had an alternative measure with a different scale (say $Z_{\theta,t,2} = \lambda_{\theta,t,2} \ln \theta_t$), we would generally obtain different parameters if we replaced $Z_{\theta,t,1}$ with $Z_{\theta,t,2}$. This alternative measure could for example result from changing the units of measurements of $Z_{\theta,t,1}$, such as using years instead of months of education.

Even though the production function parameters are not identified, there are potential interpretations that adapt to the units of measurements. For example, the identified parameter $\lambda_{\theta,t+1,1}\gamma_{2t}$ tells us the effect of a one unit increase in $\ln(I_t)$ on skills, measured in the units of $Z_{\theta,t+1,1}$ (see Section 5 for a specific example). It can also be shown that

$$\begin{aligned} F_{\ln \theta_{t+1}}(a_t + \gamma_{1t} \ln Q_\alpha(\theta_t) + \gamma_{2t} \ln i_t) &= F_{\lambda_{\theta,t+1,1} \ln \theta_{t+1}}(\lambda_{\theta,t+1,1}(a_t + \gamma_{1t} \ln Q_\alpha(\theta_t) + \gamma_{2t} \ln i_t)) \\ &= F_{Z_{\theta,t+1,1}}\left(\lambda_{\theta,t+1,1}a_t + \frac{\lambda_{\theta,t+1,1}}{\lambda_{\theta,t,1}}\gamma_{1t}Q_\alpha(Z_{\theta,t,1}) + \lambda_{\theta,t+1,1}\gamma_{2t} \ln i_t\right) \end{aligned}$$

where i_t is a fixed level of investment, $F_{\ln \theta_{t+1}}$ is the cdf of $\ln \theta_{t+1}$, and $Q_\alpha(\theta_t)$ is the α -quantile of θ_t . Since the right hand side only depends on identified parameters, similar to marginal effects in the probit model, we can identify how exogenous changes in someones investment affects her rank in the skill distribution at time $t + 1$ for a given skill quantile at time t .

As another example, first combine the production functions from two periods, write

$$\ln \theta_2 = a_1 + \gamma_{11}a_0 + \gamma_{11}\gamma_{10} \ln \theta_0 + \gamma_{11}\gamma_{20} \ln I_0 + \gamma_{21} \ln I_1,$$

and for a given investment sequence (i_0, i_1) define

$$\ln \theta_2(i_0, i_1) = a_1 + \gamma_{11}a_0 + \gamma_{11}\gamma_{10} \ln \theta_0 + \gamma_{11}\gamma_{20} \ln i_0 + \gamma_{21} \ln i_1$$

We can rewrite this equation in terms of identified features as

$$\begin{aligned}\lambda_{\theta,2,1} \ln \theta_2(i_0, i_1) &= \lambda_{\theta,2,1} a_1 + \lambda_{\theta,2,1} \gamma_{11} a_0 + \left(\frac{\lambda_{\theta,2,1}}{\lambda_{\theta,1,1}} \gamma_{11} \right) \left(\frac{\lambda_{\theta,1,1}}{\lambda_{\theta,0,1}} \gamma_{10} \right) \lambda_{\theta,0,1} \ln \theta_0 \\ &\quad + \left(\frac{\lambda_{\theta,2,1}}{\lambda_{\theta,1,1}} \gamma_{20} \right) (\lambda_{\theta,1,1} \gamma_{11}) \ln i_0 + (\gamma_{21} \lambda_{\theta,2,1}) \ln i_1\end{aligned}$$

showing that we can identify the distribution of $\lambda_{\theta,2,1} \ln \theta_2(i_0, i_1)$, which we can interpret as a counterfactual measure for a given investment sequence. Using this result, we can also identify the sequence of investment which maximizes expected log-skills. That is, consider $g(\tau_0) \in \mathbb{R}^2$ defined as

$$g(\tau_0) = \arg \max_{(i_0, i_1) \in \mathcal{I}} E [\ln \theta_2(i_0, i_1)],$$

where \mathcal{I} is a set of feasible investments. Since the maximizer is invariant to changes in scales and locations of the objective function, $g(\tau_0) = \arg \max_{(i_0, i_1) \in \mathcal{I}} E [\lambda_{\theta,2,1} \ln \theta_2(i_0, i_1)]$, showing that $g(\tau_0)$ is identified. All the identified features above are invariant to the restriction $\lambda_{\theta,t,1} = 1$, implying that $\lambda_{\theta,t,1} = 1$ is a normalization with respect to these features.

Next to primitive parameters, some counterfactuals reported in applications are not be invariant to $\lambda_{\theta,t,1} = 1$ either. As an example, suppose the production function for $\ln \theta_1$ also includes an additive interaction term, $\gamma_{30} \ln \theta_0 \ln I_0$, in which case it can be shown that

$$\lambda_{\theta,2,1} \ln \theta_2(i_0, i_1) = \zeta_0 + \zeta_1 \ln \theta_0 + \zeta_3 \ln i_0 + \zeta_4 \ln i_1 + \zeta_5 \ln \theta_0 \ln i_0$$

for identified parameters $\{\zeta_j\}_{j=1}^5$. Using $\ln \theta_0 \sim N(0, s_0^2)$ it follows that

$$E[\theta_2(i_0, i_1)^{\lambda_{\theta,2,1}}] = \exp \left(\zeta_0 + \zeta_3 \ln i_0 + \zeta_4 \ln i_1 + (\zeta_1 + \zeta_5 \ln i_0)^2 s_0^2 \right)$$

but

$$E[\theta_2(i_0, i_1)] = \exp \left(\frac{1}{\lambda_{\theta,2,1}} \left(\zeta_0 + \zeta_3 \ln i_0 + \zeta_4 \ln i_1 + (\zeta_1 + \zeta_5 \ln i_0)^2 \frac{s_0^2}{\lambda_{\theta,2,1}} \right) \right)$$

which implies that generally $\arg \max_{(i_0, i_1) \in \mathcal{I}} E [\theta_2(i_0, i_1)] \neq \arg \max_{(i_0, i_1) \in \mathcal{I}} E [\theta_2(i_0, i_1)^{\lambda_{\theta,2,1}}]$. Hence, using different scaled versions of log-skills (or different fixed values of $\lambda_{\theta,t,1}$) results

in different optimal investment sequences when studying the skill level.

Finally, consider the CES production function

$$\theta_{t+1} = (\gamma_{1t}\theta_t^{\sigma_t} + \gamma_{2t}I_t^{\sigma_t})^{1/\sigma_t}$$

The parameter vector is now $\tau_0 = (\lambda_{\theta,0,1}, \lambda_{\theta,1,1}, \lambda_{\theta,2,1}, \sigma_0, \sigma_1, \gamma_{10}, \gamma_{20}, \gamma_{11}, \gamma_{21}, s_0^2)$ and we can write the production function in terms of observables as

$$\exp(Z_{\theta,t+1,1}) = (\gamma_{1t} \exp(Z_{\theta,t,1})^{\sigma_t/\lambda_{\theta,t,1}} + \gamma_{2t} I_t^{\sigma_t})^{\lambda_{\theta,t+1,1}/\sigma_t}$$

Due to the functional form restrictions of the CES production function, it can be shown that τ_0 is point identified without any additional restrictions.

Thus, the commonly imposed scale restrictions ($\lambda_{\theta,t,1} = 1$ for all t) lead to misspecification unless the true values are all 1. Consequently, imposing this restriction yields different conclusions for different scaled versions of log-skills (or different fixed values of $\lambda_{\theta,t,1}$) irrespective of the counterfactual. Similar as in the trans-log case, many important features are identified without additional restrictions and are invariant to scaling of the measures.

These results generalize to more complicated settings as discussed in the next section.

3 Skill formation models

3.1 Model

I now discuss issues arising from normalizations in a general class of skill formation models. As before, θ_t and I_t denote skills and investment at time t , respectively. Now neither skills nor investment are directly observed and we denote the observed measurements by $Z_{\theta,t,m}$ and $Z_{I,t,m}$, respectively. Specifically, I consider the model:

$$\begin{aligned} (1) \quad \theta_{t+1} &= f(\theta_t, I_t, \delta_t, \eta_{\theta,t}) & t = 0, \dots, T-1 \\ (2) \quad Z_{\theta,t,m} &= \mu_{\theta,t,m} + \lambda_{\theta,t,m} \ln \theta_t + \varepsilon_{\theta,t,m} & t = 0, \dots, T, m = 1, 2 \\ (3) \quad Z_{I,t,m} &= \mu_{I,t,m} + \lambda_{I,t,m} \ln I_t + \varepsilon_{I,t,m} & t = 0, \dots, T-1, m = 1, 2 \end{aligned}$$

The first equation describes the production technology with a production function f that depends on skills and investment at time t , a parameter vector δ_t , and an unobserved shock $\eta_{\theta,t}$. The second and the third equation describe the measurement system for unobserved (latent) skills θ_t and unobserved investment I_t , respectively. Observed investment is a special case with $\mu_{I,t,m} = 0$, $\lambda_{I,t,m} = 1$, and $\varepsilon_{I,t,m} = 0$ for all m and t in which case $Z_{I,t,m} = \ln I_t$.

Next, I introduce two equations to allow for endogenous investment and anchoring at an adult outcome. If investment is exogenous, in the sense that $\eta_{\theta,t}$ is independent of I_t , then these equations are not needed for the main identification results. That is, let

$$(4) \quad \ln I_t = \beta_{0t} + \beta_{1t} \ln \theta_t + \beta_{2t} \ln Y_t + \eta_{I,t} \quad t = 0, \dots, T-1$$

$$(5) \quad Q = \rho_0 + \rho_1 \ln \theta_T + \eta_Q$$

Here Y_t is parental income (or another exogenous variable that affects investment) and Q is an adult outcome, such as earnings or education. An adult outcome does not necessarily have to be available and we can simply use a skill measure in period T in its place.

In summary, the observed variables are income $\{Y_t\}_{t=0}^{T-1}$, the measures $\{Z_{\theta,t,m}\}_{t=0,\dots,T-1,m=1,2}$ and $\{Z_{I,t,m}\}_{t=0,\dots,T-1,m=1,2}$, and the adult outcome Q , but we neither observe skills $\{\theta_t\}_{t=0}^T$ nor investment $\{I_t\}_{t=0}^{T-1}$. We also do not observe any of the errors/shocks in the five equations. The parameters are $\{\mu_{\theta,t,m}, \lambda_{\theta,t,m}\}_{t=0,\dots,T-1,m=1,2}$, $\{\mu_{I,t,m}, \lambda_{I,t,m}\}_{t=0,\dots,T-1,m=1,2}$, $\{\delta_t\}_{t=0}^{T-1}$, $\{\beta_{0t}, \beta_{1t}, \beta_{2t}\}_{t=0}^{T-1}$, and (ρ_0, ρ_1) .

In the following analysis, I consider the two most commonly used forms for the production technology in the empirical literature, namely the trans-log production function with

$$(6) \quad \ln \theta_{t+1} = a_t + \gamma_{1t} \ln \theta_t + \gamma_{2t} \ln I_t + \gamma_{3t} \ln \theta_t \ln I_t + \eta_{\theta,t}$$

and parameter vector $\delta_t = (a_t, \gamma_{1t}, \gamma_{2t}, \gamma_{3t})$ and the CES production function with

$$(7) \quad \theta_{t+1} = (\gamma_{1t} \theta_t^{\sigma_t} + \gamma_{2t} I_t^{\sigma_t})^{\psi_t / \sigma_t} \exp(\eta_{\theta,t})$$

and parameter vector $\delta_t = (\gamma_{1t}, \gamma_{2t}, \sigma_t, \psi_t)$. When $\gamma_{3t} = 0$, the trans-log reduces to the Cobb-Douglas production function.

I now state several additional assumptions that are common in the literature.

Assumption 1.

- (a) $\{\{\varepsilon_{\theta,t,m}\}_{t=0,\dots,T,m=1,2}, \{\varepsilon_{I,t,m}\}_{t=0,\dots,T-1,m=1,2}, \eta_Q\}$ are jointly independent and independent of $\{\{\theta_t\}_{t=0}^T, \{I_t\}_{t=0}^{T-1}\}$ conditional on $\{Y_t\}_{t=0}^{T-1}$.
- (b) All random variables have bounded first and second moments.
- (c) $E[\varepsilon_{\theta,t,m}] = E[\varepsilon_{I,t,m}] = E[\varepsilon_Q] = 0$ for all t and m .
- (d) $\lambda_{\theta,t,m}, \lambda_{I,t,m} \neq 0$ for all t and m .
- (e) For all $t \in \{0, \dots, T\}$, $\text{cov}(\ln \theta_t, \ln I_s) \neq 0$ for some $s \in \{0, \dots, T-1\}$ or $\text{cov}(\ln \theta_t, \ln \theta_s) \neq 0$ for some $s \in \{0, \dots, T\} \setminus t$. For all $t \in \{0, 1, \dots, T-1\}$, $\text{cov}(\ln I_t, \ln \theta_s) \neq 0$ for some $s \in \{0, \dots, T\}$ or $\text{cov}(\ln I_t, \ln I_s) \neq 0$ for some $s \in \{0, \dots, T-1\} \setminus t$.
- (f) For all t and m the real zeros of the characteristic functions of $\varepsilon_{\theta,t,m}$ are isolated and are distinct from those of its derivatives. Identical conditions hold for the characteristic functions of $\varepsilon_{I,t,m}$ and η_Q .
- (g) The support of (θ_t, I_t, Y_t) includes an open ball in \mathbb{R}^3 for all t .
- (h) $E[\eta_{I,t} \mid \theta_t, Y_t] = 0$ and $E[\eta_{\theta,t} \mid \theta_t, \eta_{I,t}, Y_t] = \kappa_t \eta_{I,t}$ for all t .

Part (a) imposes common independence assumptions on the measurement errors. Importantly, I_t and θ_t are not independent and I_t may be endogenous and contemporaneously correlated with $\eta_{\theta,t}$. Part (b) is a standard restriction, part (c) is needed because all measurement equations contain an intercept, and part (d) ensures that the skills actually affect the measures. Part (e) requires that skills and investment are correlated in some time periods. Sufficient conditions are that $\text{cov}(\ln \theta_{t+1}, \ln \theta_t) \neq 0$ and $\text{cov}(\ln I_{t+1}, \ln I_t) \neq 0$ for all t . Under parts (a) and (d) zero covariances of the latent variables are identified because, for example, $\text{cov}(\ln \theta_t, \ln \theta_s) = 0$ if and only if $\text{cov}(Z_{\theta,t,1}, Z_{\theta,s,1}) = 0$. Notice that I only require two measures in each period. One can drop part (e) by assuming that three measures are available. Part (f) contains weak regularity conditions needed for nonparametric identification of the

distributions of skills and investment and that hold for most common distributions. Part (g) is a mild support condition that ensures sufficient variation of (θ_t, I_t, Y_t) . Part (h) implies that Y_t can serve as an instrument with identification based on a control function argument, as in Attanasio et al. (2020). Linearity of the conditional mean function can be relaxed to allow for more flexible functional forms. Exogenous investment is a special case with $\kappa_t = 0$.

Under parts (a)–(f) of Assumption 1 we get the following result.

Lemma 1. *Suppose that parts (a)–(f) of Assumption 1 hold. Then the joint distribution of*

$$(\{\mu_{\theta,t,m} + \lambda_{\theta,t,m} \ln \theta_t\}_{t=0,\dots,T,m=1,2}, \{\mu_{I,t,m} + \lambda_{I,t,m} \ln I_t\}_{t=0,\dots,T-1,m=1,2}, \rho_0 + \rho_1 \ln \theta_T)$$

is point identified conditional on $\{Y_t\}_{t=0}^{T-1}$.

The proof follows from an extension of Kotlarski’s Lemma (Evdokimov and White 2012). Lemma 1 shows that under Assumption 1, the distribution of linear combinations of log-skills and investments is identified. However, Assumption 1 does not imply identification of the parameters in equations (1)–(3), such as δ_t . Below I discuss additional assumptions, which have been used in the literature, to achieve point identification of two sets of parameters: (i) the primitive parameters of the model (1)–(5): $\{\mu_{\theta,t,m}, \lambda_{\theta,t,m}\}_{t=0,\dots,T,m=1,2}$, $\{\mu_{I,t,m}, \lambda_{I,t,m}\}_{t=0,\dots,T-1,m=1,2}$, $\{\delta_t\}_{t=0}^{T-1}$, $\{\beta_{0t}, \beta_{1t}, \beta_{2t}\}_{t=0}^{T-1}$, and (ρ_0, ρ_1) and (ii) “policy relevant” parameters, such as how changes in investment or income affect Q . I consider various combinations of assumptions and I discuss in what sense restrictions are normalizations.

There are two separate issues concerning identification and normalizations in this model. First, the previous literature has focused on sufficient conditions for point identification, which includes scale and location restrictions. However, it is unclear whether these restrictions are necessary. If they instead restrict identified parameters, the model may be misspecified and estimators are generally inconsistent. Second, even if the restrictions simply select an element of the identified set, it is important to understand which features are invariant to arbitrary scale and location restrictions. Although the implications in the CES case are more interesting, I start with the more transparent trans-log case where only the second issue arises. As pointed out before, whether or not a restriction is a normalization depends on the model, and parameters may be invariant in some settings, but not in others.

3.2 Trans-log production function

In this section I consider the trans-log production function given in equation (6). I first introduce additional assumptions that are commonly used in the literature.

Assumption 2. $\lambda_{\theta,0,1} = 1$ and $\mu_{\theta,0,1} = 0$.

Assumption 3.

(a) $\lambda_{\theta,t,1} = \lambda_{\theta,t+1,1}$ and $\mu_{\theta,t,1} = \mu_{\theta,t+1,1}$ for all $t = 0, \dots, T-1$

(b) $a_t = 0$ and $\gamma_{1t} + \gamma_{2t} + \gamma_{3t} = 1$ for all $t = 0, \dots, T-1$.

Assumption 4.

(a) $\lambda_{I,t,1} = 1$ and $\mu_{I,t,1} = 0$ for all $t = 0, \dots, T-1$.

(b) $\beta_{0t} = 0$ and $\beta_{1t} + \beta_{2t} = 1$ for all $t = 0, \dots, T-1$.

Assumption 2 is usually thought of as a normalization, which is commonly imposed since log-skills are only identified up to scale and location. Here, I impose the restrictions on the first measure, which is without loss of generality. Instead of fixing the intercept and the slope coefficient in equation (2) for $t = 0$, one could set $\rho_0 = 0$ and $\rho_1 = 1$ and thus “anchor” the skills at Q . Assumption 2 anchors the skills at $Z_{\theta,0,1}$, but analogous issues discussed here arise with anchoring at Q (see Section 3.2.4). Without such an assumption, the parameters are not point identified. Assumption 3(a) states that the skill measures are age-invariant (using the terminology of Agostinelli and Wiswall (2024) – see their Definition 1 and footnote 10). Assumption 3(b) imposes restrictions on the technology, which Agostinelli and Wiswall (2024) refer to as a known scale and location assumption in a more general context. Assumption 4(a) says that an investment measure is age-invariant. Assumption 4(b) states constant return to scale in equation (4), which is a strong assumption and used for point identification without age-invariant investment measures. If investment is observed (i.e. $Z_{I,t,m} = \ln I_t$), Assumption 4(a) is automatically satisfied.

I now characterize the identified set of the primitive parameters under Assumption 1 only. I then discuss point identification under different combinations of Assumptions 1–4 and show

that several policy relevant parameters are invariant to the restrictions in Assumption 2 and are in fact point identified under Assumption 1. Finally, I illustrate why Assumption 2 is generally not a normalization for the primitive as well as some policy relevant parameters.

3.2.1 Identification

Define $\tilde{I}_t = \exp(\mu_{I,t,1})I_t^{\lambda_{I,t,1}}$ and $\tilde{\theta}_t = \exp(\mu_{\theta,t,1})\theta_t^{\lambda_{\theta,t,1}}$ so that

$$\ln \tilde{\theta}_t = \mu_{\theta,t,1} + \lambda_{\theta,t,1} \ln \theta_t \quad \text{and} \quad \ln \theta_t = \frac{\ln \tilde{\theta}_t - \mu_{\theta,t,1}}{\lambda_{\theta,t,1}}.$$

We can then rewrite the production function in terms of $\tilde{\theta}_t$ and \tilde{I}_t because

$$\frac{\ln \tilde{\theta}_{t+1} - \mu_{\theta,t+1,1}}{\lambda_{\theta,t+1,1}} = a_t + \gamma_{1t} \frac{\ln \tilde{\theta}_t - \mu_{\theta,t,1}}{\lambda_{\theta,t,1}} + \gamma_{2t} \frac{\ln \tilde{I}_t - \mu_{I,t,1}}{\lambda_{I,t,1}} + \gamma_{3t} \frac{\ln \tilde{\theta}_t - \mu_{\theta,t,1}}{\lambda_{\theta,t,1}} \frac{\ln \tilde{I}_t - \mu_{I,t,1}}{\lambda_{I,t,1}} + \eta_{\theta,t}.$$

After rearranging, we can then rewrite equations (1)–(5) as

$$(8) \quad \ln \tilde{\theta}_{t+1} = \tilde{a}_t + \tilde{\gamma}_{1t} \ln \tilde{\theta}_t + \tilde{\gamma}_{2t} \ln \tilde{I}_t + \tilde{\gamma}_{3t} \ln \tilde{\theta}_t \ln \tilde{I}_t + \tilde{\eta}_{\theta,t} \quad t = 0, \dots, T-1$$

$$(9) \quad Z_{\theta,t,m} = \tilde{\mu}_{\theta,t,m} + \tilde{\lambda}_{\theta,t,m} \ln \tilde{\theta}_t + \varepsilon_{\theta,t,m} \quad t = 0, \dots, T, m = 1, 2$$

$$(10) \quad Z_{I,t,m} = \tilde{\mu}_{I,t,m} + \tilde{\lambda}_{I,t,m} \ln \tilde{I}_t + \varepsilon_{I,t,m} \quad t = 0, \dots, T-1, m = 1, 2$$

$$(11) \quad \ln \tilde{I}_t = \tilde{\beta}_{0t} + \tilde{\beta}_{1t} \ln \tilde{\theta}_t + \tilde{\beta}_{2t} \ln Y_t + \tilde{\eta}_{I,t} \quad t = 0, \dots, T-1$$

$$(12) \quad Q = \tilde{\rho}_0 + \tilde{\rho}_1 \ln \tilde{\theta}_T + \eta_Q$$

where

$$\tilde{\gamma}_{1t} = \frac{\lambda_{\theta,t+1,1}}{\lambda_{\theta,t,1}} \left(\gamma_{1t} - \frac{\mu_{I,t,1}}{\lambda_{I,t,1}} \gamma_{3t} \right), \quad \tilde{\gamma}_{2t} = \frac{\lambda_{\theta,t+1,1}}{\lambda_{I,t,1}} \left(\gamma_{2t} - \frac{\mu_{\theta,t,1}}{\lambda_{\theta,t,1}} \gamma_{3t} \right), \quad \tilde{\gamma}_{3t} = \frac{\lambda_{\theta,t+1,1}}{\lambda_{\theta,t,1} \lambda_{I,t,1}} \gamma_{3t}$$

and expressions for all other parameters in (8)–(12) are provided in Appendix A.1. Importantly, by construction, $\tilde{\mu}_{\theta,t,1} = \tilde{\mu}_{I,t,1} = 0$ and $\tilde{\lambda}_{\theta,t,1} = \tilde{\lambda}_{I,t,1} = 1$ and Lemma 1 implies that the joint distribution of $(\{\ln \tilde{\theta}_t\}_{t=0}^T, \{\ln \tilde{I}_t\}_{t=0}^T)$ is point identified conditional on $\{Y_t\}_{t=0}^{T-1}$, which then yields identification of all parameters in (8)–(12). These parameters are functions of the primitive parameters in (1)–(5). The next theorem, which characterizes the identified

set, states that the identified set consists of all primitive parameters that imply the same values in (8)–(12) as the true parameters, which is very similar to the probit model.

Theorem 1. *Suppose Assumption 1 holds.*

1. *The identified set of $\{\mu_{\theta,t,m}, \lambda_{\theta,t,m}\}_{t=0,\dots,T,m=1,2}$, $\{\mu_{I,t,m}, \lambda_{I,t,m}\}_{t=0,\dots,T-1,m=1,2}$, (ρ_0, ρ_1) , $\{\beta_{0t}, \beta_{1t}, \beta_{2t}\}_{t=0}^{T-1}$, $\{a_t, \gamma_{1t}, \gamma_{2t}, \gamma_{3t}\}_{t=0}^{T-1}$ consists of all vectors that yield the same values of $\{\tilde{\mu}_{\theta,t,m}, \tilde{\lambda}_{\theta,t,m}\}_{t=0,\dots,T,m=1,2}$, $\{\tilde{\mu}_{I,t,m}, \tilde{\lambda}_{I,t,m}\}_{t=0,\dots,T-1,m=1,2}$, $(\tilde{\rho}_0, \tilde{\rho}_1)$, $\{\tilde{\beta}_{0t}, \tilde{\beta}_{1t}, \tilde{\beta}_{2t}\}_{t=0}^{T-1}$, and $\{\tilde{a}_t, \tilde{\gamma}_{1t}, \tilde{\gamma}_{2t}, \tilde{\gamma}_{3t}\}_{t=0}^{T-1}$ as the true parameter vectors.*
2. *Let $\{\bar{\mu}_{\theta,t,1}, \bar{\lambda}_{\theta,t,1}\}_{t=0}^T$, $\{\bar{\mu}_{I,t,1}, \bar{\lambda}_{I,t,1}\}_{t=0}^{T-1}$, be fixed constants with $\bar{\lambda}_{\theta,t,1}, \bar{\lambda}_{I,t,1} \neq 0$ for all t . If in addition $\{\mu_{\theta,t,1}, \lambda_{\theta,t,1}\}_{t=0}^T = \{\bar{\mu}_{\theta,t,1}, \bar{\lambda}_{\theta,t,1}\}_{t=0}^T$ and $\{\mu_{I,t,1}, \lambda_{I,t,1}\}_{t=0}^{T-1} = \{\bar{\mu}_{I,t,1}, \bar{\lambda}_{I,t,1}\}_{t=0}^{T-1}$, then the identified set is a singleton.*

Part 2 of the theorem shows that the parameters are indeed not point identified under Assumption 1 only and that the sources of underidentification are the ambiguous scales and locations of skills and investments. For example, without additional assumptions, equations (1)–(5) and (8)–(12) are observationally equivalent, and we cannot distinguish between the skills θ_t and $\tilde{\theta}_t$ and the corresponding production functions. Even if investment was observed (and $\mu_{I,t,1} = 0$ for all t), we can then only identify $(\lambda_{\theta,t+1,m}/\lambda_{\theta,t,m})\gamma_{1t}$, but not $\lambda_{\theta,t,m}$ and γ_{1t} separately. Hence, we cannot distinguish between changes in the quality of the measurements $(\lambda_{\theta,t+1,m}/\lambda_{\theta,t,m})$ and changes in the technology (γ_{1t}). For example, suppose $Z_{\theta,t,m}$ are test scores. We then cannot distinguish between all children getting smarter or tests becoming easier. Similarly, we can at best identify γ_{2t} up to scale, even with observed investment.

The following corollary states that all parameters are point identified under additional assumptions. These results are an extension of those in Agostinelli and Wiswall (2024), who assume that investment is exogenous (in the sense that it is uncorrelated with $\eta_{\theta,t}$).

Corollary 1. *Suppose Assumptions 1 and 2 hold. Suppose either Assumption 3(a) or Assumption 3(b) holds. Suppose either Assumption 4(a) or Assumption 4(b) holds. Then all parameters are point identified.*

The corollary also immediately implies that Assumptions 3(a) and 3(b) together impose additional testable restrictions, which is one of the main contributions of Agostinelli and

Wiswall (2016a, 2016b, 2024). Contrarily, as shown in Theorem A.1 in the appendix and illustrated in examples below, if the model is correctly specified and Assumption 1 holds, then there always exist sets of parameters which are consistent with the data and satisfy Assumptions 1, 2, either 3(a) or 3(b), and either 4(a) or 4(b). These different sets of assumptions therefore impose no additional restrictions on the distribution of observables. While different sets of assumptions yield point identification and are observationally equivalent, the estimated primitive parameters are usually quite different, as illustrated in Section 3.2.3.

3.2.2 Invariant parameters

I now show that many interesting features are point identified under Assumption 1 only because they can all be rewritten in terms of the identified parameters in (8)–(12), as in Section 2.2. They include summaries of the productions functions and effects of investment and income on skills and adult outcomes. These features do not constitute an exhaustive list and there may be many others. To state the formal results, recall that $Q_\alpha(\theta_t)$ is the α quantile of the skill distribution at time t and $F_{\ln(\theta_{t+1})}(\cdot)$ is the cdf of log-skills at time $t+1$. Define $s_{1t}(\alpha_1, \alpha_2, \alpha_3) = a_t + \gamma_{1t} \ln Q_{\alpha_1}(\theta_t) + \gamma_{2t} \ln Q_{\alpha_2}(I_t) + \gamma_{3t} \ln Q_{\alpha_1}(\theta_t) \ln Q_{\alpha_2}(I_t) + Q_{\alpha_3}(\eta_{\theta,t})$ which are the log-skills in period $t+1$ for specific quantiles of the inputs in period t . Similarly, let $s_{2t}(\alpha_1, \alpha_2, \alpha_3, y) = a_t + \gamma_{1t} \ln Q_{\alpha_1}(\theta_t) + \gamma_{2t} \ln I_t(y) + \gamma_{3t} \ln Q_{\alpha_1}(\theta_t) \ln I_t(y) + Q_{\alpha_2}(\eta_{\theta,t})$, where $\ln I_t(y) = \beta_{0t} + \beta_{1t} \ln Q_{\alpha_1}(\theta_t) + \beta_{2t} \ln y + Q_{\alpha_3}(\eta_{I,t})$, which are log-skills and log-investment, that depend on a specific exogenously set value of Y .

Theorem 2. *Suppose Assumption 1 holds.*

1. $F_{\ln \theta_{t+1}}(s(\alpha_1, \alpha_2, \alpha_3))$ and $\mu_{\theta,t+1,m} + \lambda_{\theta,t+1,m} s_{1t}(\alpha_1, \alpha_2, \alpha_3) + Q_{\alpha_4}(\varepsilon_{\theta,t+1,m})$ are point identified for all $\{\alpha_j\}_{j=1}^4 \in (0, 1)$. $\frac{\partial \lambda_{\theta,t+1,m} \ln \theta_{t+1}}{\partial \lambda_{\theta,t,m'} \ln \theta_t} \big|_{I_t=Q_{\alpha_2}(I_t)}$ and $\frac{\partial \lambda_{\theta,t+1,m} \ln \theta_{t+1}}{\partial \lambda_{I,t,m'} \ln I_t} \big|_{\theta_t=Q_{\alpha_1}(\theta_t)}$ are point identified for all m, m' .
2. $F_{\ln \theta_{t+1}}(s_{2t}(\alpha_1, \alpha_2, \alpha_3, y))$ and $\mu_{\theta,t+1,m} + \lambda_{\theta,t+1,m} s_{2t}(\alpha_1, \alpha_2, \alpha_3, y) + Q_{\alpha_4}(\varepsilon_{\theta,t+1,m})$ are point identified for all $\{\alpha_j\}_{j=1}^4 \in (0, 1)$.
3. $P(Q \leq q \mid \theta_s = Q_{\alpha_1}(\theta_s), \{I_t = Q_{\alpha_2}(I_t)\}_{t=0}^{T-1}, \{\eta_{\theta,t} = Q_{\alpha_3}(\eta_{\theta,t})\}_{t=s}^{T-1})$ is point identified for all $\alpha_1, \{\alpha_{2t}, \alpha_{3t}\}_{t=s}^{T-1} \in (0, 1)$.

4. $P(Q \leq q \mid \theta_s = Q_\alpha(\theta_s), \{Y_t = y_t\}_{t=s}^{T-1})$ is point identified for all $\alpha \in (0, 1)$.

5. Suppose Assumption 3(a) also holds. Then $\gamma_{1t} + \gamma_{3t} \ln Q_\alpha(I_t)$ is point identified for all α and $P(\gamma_{1t} + \gamma_{3t} \ln I_t \leq q)$ is point identified for all $q \in \mathbb{R}$.

The function $F_{\ln \theta_{t+1}}(a_t + \gamma_{1t} \ln Q_{\alpha_1}(\theta_t) + \gamma_{2t} \ln Q_{\alpha_2}(I_t) + \gamma_{3t} \ln Q_{\alpha_1}(\theta_t) \ln Q_{\alpha_2}(I_t) + Q_{\alpha_3}(\eta_{\theta,t}))$ measures how changes in skills and investment changes the relative standing in the skill distribution. For example, consider an individual with $\theta_t = Q_{0.1}(\theta_t)$, meaning that the person is at lowest 10% of the skill distribution at time t . Then, given investment $I_t = Q_{0.25}(I_t)$ and a median production function shock, $\eta_{\theta,t} = Q_{0.5}(\eta_{\theta,t})$, $F_{\ln \theta_{t+1}}(s(0.1, 0.25, 0.5))$ tells us the relative rank (or the quantile) in the skill distribution at time $t + 1$. We can then for example vary the investment quantile and analyze how future skill ranks are affected. This feature is identified because I show in the appendix that

$$\begin{aligned} & F_{\ln \theta_{t+1}}(a_t + \gamma_{1t} Q_{\alpha_1} \ln(\theta_t) + \gamma_{2t} Q_{\alpha_2}(\ln I_t) + \gamma_{3t} Q_{\alpha_1} \ln(\theta_t) Q_{\alpha_2}(\ln I_t) + Q_{\alpha_3}(\eta_{\theta,t})) \\ &= F_{\ln \tilde{\theta}_{t+1}}(\tilde{a}_t + \tilde{\gamma}_{1t} \ln Q_{\alpha_1}(\tilde{\theta}_t) + \tilde{\gamma}_{2t} \ln Q_{\alpha_2}(\tilde{I}_t) + \tilde{\gamma}_{3t} \ln Q_{\alpha_1}(\tilde{\theta}_t) \ln Q_{\alpha_2}(\tilde{I}_t) + Q_{\alpha_3}(\tilde{\eta}_{\theta,t})) \end{aligned}$$

Thus, one can estimate the model based on (8)–(12) and calculate the feature using $\tilde{\theta}_t$ instead of θ_t . The estimand then corresponds to the true effect, even if Assumptions 2–4 do not hold.

Once we know the rank at time $t + 1$ and fix investment and production shock quantiles in that period, we can identify the skill rank at time $t + 2$. Using recursive arguments, we can identify the relative rank in period T , given investment and production shock quantiles in all period and a skill quantile in period 0. We could then make statements such as: “A person at lowest 10% of the skill distribution in period 0 would end up at the 30% quantile of the original skill distribution in the final period with a particular investment strategy and median production function shocks.” These statements would allow comparisons of investment strategies, assessing heterogeneous effects, and choosing optimal investments.

Instead of considering ranks of skills, one can interpret the effects in the units of any of the measures. For example, changing investment from $Q_{\alpha_2}(I_t)$ to $Q_{\alpha'_2}(I_t)$ changes skills at time $t + 1$ such that skill measure m changes by $\lambda_{\theta,t+1,m}(s_1(\alpha_1, \alpha'_2, \alpha_3) - s_1(\alpha_1, \alpha_2, \alpha_3))$ (holding everything else equal). We can also identify that a change in investment, which

corresponds to a 1 unit increase in investment measure m' at time t , affects skills at time $t + 1$ in a way that skill measure m changes by $\frac{\partial \lambda_{\theta, t+1, m} \ln \theta_{t+1}}{\partial \lambda_{I, t, m'} \ln I_t} |_{\theta_t = Q_{\alpha_1}(\theta_t)}$ (see Section 5 for specific examples).

The second part is very similar, but it also considers exogenous changes in income.

Instead of considering the skill rank in the final period, we could also look at the distribution of the adult outcome Q (or, alternatively, one of the skill measures in the final period). Notice that it is important to condition on the production function shocks, because investment is endogenous. We can either fix a quantile or average over its marginal distribution since we obtain identification for all quantiles. The fourth part focuses on the effect of income on skills. Again, we can point identify averages and features of the distribution such as $\int E(Q | \theta_s = \theta, \{Y_t = y_t\}_{t=s}^{T-1}) f_{\theta_s}(\theta) d\theta$ (which differs from $E(Q | \{Y_t = y_t\}_{t=s}^{T-1})$ due to a potential dependence between $\{Y_t = y_t\}_{t=s}^{T-1}$ and skills). Also notice that

$$\int E(Q | \theta_s = \theta, \{Y_t = y_t\}_{t=s}^{T-1}) f_{\theta_s}(\theta) d\theta = \rho_0 + \rho_1 \int E(\ln \theta_T | \theta_s = \theta, \{Y_t = y_t\}_{t=s}^{T-1}) f_{\theta_s}(\theta) d\theta$$

Thus, we can identify the sequence $\{Y_t = y_t\}_{t=s}^{T-1}$ that maximizes the conditional expected value of $\ln \theta_T$. For this sequence, we do not necessarily need to observe an adult outcome because we can instead use one of the skill measures in period T . To identify these features, one only has to identify the joint distribution of $(Q, Y_1, \dots, Y_{T-1}, \tilde{\theta}_s)$. For example, when $s = 0$, we do not require any skill measures in periods $1, 2, \dots, T$.³ Importantly objects such as $\int E(\theta_T | \theta_s = \theta, \{Y_t = y_t\}_{t=s}^{T-1}) f_{\theta_s}(\theta) d\theta$ are not point identified without the scale and location restrictions and are sensitive to the specific values used (see Example 2 below).

Remark 1. To summarize the production technology, Del Bono, Kinsler, and Pavan (2022) show that standardizing skills can lead to features that are invariant to scale and location

³Part 4 of Theorem 2 also implies identification of

$$\int \int \left(\frac{\partial}{\partial y_s} E(Q | \theta_s = \theta, \{Y_t = y_t\}_{t=s}^{T-1}) \right) f_{\theta_s, Y_s, \dots, Y_{T-1}}(\theta, y_s, \dots, y_{T-1}) d\theta dy_s \cdots dy_{T-1}$$

which Del Bono et al. (2022) refer to as “anchored treatment effects”. Del Bono et al. (2022) show that these effects are identified without age-invariant measures. My results show that one in fact does not even need any measures of investments or measures of skills in periods $s+1, 2, \dots, T$ (and therefore also no assumptions on the measurement systems, including age-invariance and independence). In addition, my arguments are not specific to the trans-log production function and carry over to other settings.

restrictions and age-invariance. In particular, they show identification of the distribution of $\left(\frac{\partial \ln \theta_{t+1}}{\partial \ln I_t}\right) / std(\ln \theta_{t+1})$. One can then make statements such as “increasing investment by 1%, increases log-skills by $x \times std(\ln \theta_{t+1})$ ”. Part 1 of Theorem 2 offers an alternative interpretation in terms of ranks or units of the measures.

Remark 2. Instead of using a two-step approach, where the distribution of a linear combination of skills and investment is identified first, Agostinelli and Wiswall (2024) substitute measures into the production function equation and use IV arguments with exogenous investment (i.e. I_t and $\eta_{\theta,t}$ are independent in (4)). Aside from exogenous investment, there are no substantial differences between the required assumptions, as both approaches are based on the joint distribution of the measures. My main contributions are to study the roles of the scale and location restrictions on the parameters of the model, to show which restrictions select an element of the identified set, and to provide features that are invariant to them and are identified without age-invariant measures and restrictions on the production function.

3.2.3 Non-invariant parameters

As shown above, under Assumption 1, there exist different sets of observationally equivalent parameters that cannot be distinguished using the data. I now illustrate with two examples that the resulting primitive parameters and counterfactuals might differ considerably.

Example 1. For simplicity, I assume that investment is observed and exogenous. In this case, Assumption 4(a) holds. I first consider a data generating process (DGP) satisfying Assumptions 1, 3(a), and 3(b), but not Assumption 2. I then construct two alternative sets of parameters, both of which are observationally equivalent to the original DGP. One of these sets of parameters satisfies Assumptions 1, 2, and 3(a) and the other set satisfies Assumptions 1, 2, and 3(b). Specifically, first assume that

$$\begin{aligned}\ln \theta_{t+1} &= \frac{1}{2} \ln \theta_t + \frac{1}{2} \ln I_t + \eta_{\theta,t} \\ \tilde{Z}_{\theta,t,1} &= \ln \theta_t + \tilde{\varepsilon}_{\theta,t,1}\end{aligned}$$

Here, I set $a_t = \mu_{\theta,t,m} = \rho_0 = 0$ and focus on the scale restriction in Assumption 2 only. For brevity, I omit equations for investment, additional measurements, and the adult outcome.

Measures often do not have a natural scale. For example, we could divide all test scores by its standard deviation or we could measure education in months rather than years. One would then hope that changing the units of measurement does not affect the economic interpretation of the results. As a specific example, suppose we estimate the model using a scaled version of the measures, namely $Z_{\theta,t,1} = 12\tilde{Z}_{\theta,t,1}$. Then for all t

$$\begin{aligned}\ln \theta_{t+1} &= \frac{1}{2} \ln \theta_t + \frac{1}{2} \ln I_t + \eta_{\theta,t} \\ Z_{\theta,t,1} &= 12 \ln \theta_t + \varepsilon_{\theta,t,1}\end{aligned}$$

where $\varepsilon_{\theta,t,1} = 12\tilde{\varepsilon}_{\theta,t,1}$. Without knowing the true DGP, one would typically impose assumptions that yield point identification when estimating the model. By Corollary 1, one could use either Assumption 3(a) or 3(b) next to Assumptions 1 and 2.

I first construct parameters satisfying Assumptions 1, 2, and 3(b) using Theorem A.1 in the appendix, which implies that there are $\{\bar{\theta}_t\}_{t=0}^T$ such that

$$\begin{aligned}\ln \bar{\theta}_{t+1} &= \bar{\gamma}_{1t} \ln \bar{\theta}_t + (1 - \bar{\gamma}_{1t}) \ln I_t + \bar{\eta}_{\theta,t} \\ Z_{\theta,t,1} &= \bar{\lambda}_{\theta,t,1} \ln \bar{\theta}_t + \varepsilon_{\theta,t,1}\end{aligned}$$

where

$$\begin{aligned}(\bar{\gamma}_{10}, \bar{\gamma}_{11}, \bar{\gamma}_{12}, \bar{\gamma}_{13}, \bar{\gamma}_{14}, \dots) &= (0.077, 0.351, 0.435, 0.470, 0.485, \dots) \\ (\bar{\lambda}_{\theta,0,1}, \bar{\lambda}_{\theta,1,1}, \bar{\lambda}_{\theta,2,1}, \bar{\lambda}_{\theta,3,1}, \bar{\lambda}_{\theta,4,1}, \dots) &= (1, 6.5, 9.25, 10.625, 11.3125, \dots)\end{aligned}$$

Moreover $\bar{\gamma}_{1t} \rightarrow \gamma_{1t} = 1/2$ and $\bar{\lambda}_{\theta,t,1} \rightarrow \lambda_{\theta,t,1} = 12$ as $t \rightarrow \infty$. Imposing the restriction $\bar{\lambda}_{\theta,0,1} = 1$ means that we estimate a model with alternative skills $\ln \bar{\theta}_0 = 12 \ln \theta_0$ and that we obtain different parameters and skill distributions. Although these two models suggest very different dynamics, they generate identical measurements. Since the true DGP is unknown, the primitive parameters are therefore hard to interpret. For example the coefficient in front of investment in the original model might be interpreted as: “increasing investment by 1%, increases skills in the next period by 0.5% and the effect is the same for all time periods” (see

e.g. Agostinelli and Wiswall (2024) for these interpretations). Contrarily, one might interpret the coefficients in the alternative model with Assumption 2 as: “increasing investment by 1% in period 0, increases skills in period 1 by 0.923% and increasing investment by 1% in period 4, increases skills in period 5 by 0.515%”, suggesting that investment is more beneficial in early periods. To obtain interpretable parameters, notice that a one unit increase in $\ln(I_t)$ leads to an increase in skills that corresponds to a $\lambda_{\theta,t+1,1}\gamma_{2t} = 6$ units increase in $Z_{\theta,t+1,1}$ or a $1/2$ unit increase in $\tilde{Z}_{\theta,t+1,1}$. This effect does not depend on the set of restrictions and its interpretation adapts to the units of measurement.

The consequences of imposing Assumptions 1, 2 and 3(a) are discussed in Section 2.2, which shows that the production function becomes $\ln \tilde{\theta}_{t+1} = \frac{1}{2} \ln \tilde{\theta}_t + 6 \ln I_t + \tilde{\eta}_{\theta,t}$.

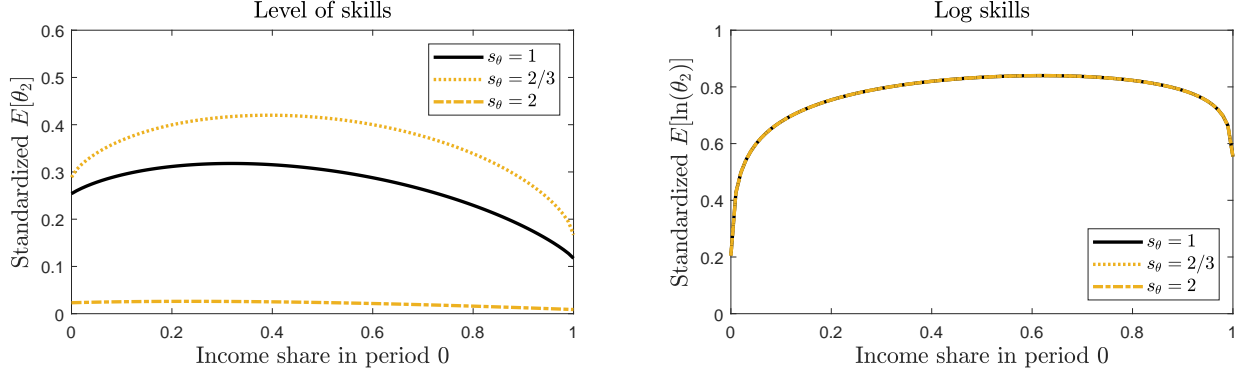
Example 2. I now use a more involved DGP from Section 4, which is based on the simulations of Attanasio et al. (2020).⁴ In this setup $T = 2$ and Assumptions 1, 2, 3(a), and 4(a) hold. Importantly, one skill and one investment measure have loadings of 1. Now suppose Y_t represents income, we can exogenously increase the sum of income of each individual in periods 0 and 1 by two standard deviations, and we want to distribute that income optimally across the two periods. Denote the skills in period 2 as a function of income by $\theta_2(Y_0 + wx, Y_1 + (1 - w)x)$, where Y_0 and Y_1 are the original incomes, x is the additional amount to be distributed, and w is the share invested in period 0. The left panel of Figure 1 shows

$$\frac{E[\theta_2(Y_0 + wx, Y_1 + (1 - w)x)] - E[\theta_2(Y_0, Y_1)]}{sd(\theta_2(Y_0, Y_1))}$$

as a function of w . That is, the y -axis shows the increase in the mean measured in standard deviations. The black line shows the effect using the true parameters, leading to an optimal income share of around 32% in period 0. Next, I multiply all skill measures $Z_{\theta,t,1}$ by s_θ and reestimate the model. Scaling the measures has identical effects to imposing $\lambda_{\theta,0,1} = 1$, when the data is generated with a loading of s_θ . The implied effects of increasing income for $s_\theta = 2/3$ and $s_\theta = 2$ can be seen in the left panel of Figure 1 as well. Depending on the scale, we obtain inefficient optimal investment choices or erroneous benefits of investment. Intuitively, the reason is that we now maximize $E[\tilde{\theta}_t(Y_0 + wx, Y_1 + (1 - w)x)] = E[\theta_t(Y_0 +$

⁴In Section 4 I use a CES production function, as do Attanasio et al. (2020), but for this numerical examples I use a trans-log production function that leads to similar observed data.

Figure 1: Mean response for different weights



Notes: This figure shows changes in standardized means of the level (left panel) and the log (right panel) of skills for different income transfers and different scales of log-skills.

$w x, Y_1 + (1 - w)x)^{s_\theta}]$ which is not invariant to s_θ . While such counterfactuals could be relevant if a welfare function is a function of the skill level, they are not identified without fixing the scales and locations and then depend on the units of measurements of the data.

The right panel contains results for the log-skills rather than the level. As discussed below Theorem 2, in this case, the optimal income sequence does not depend on the scale of the measures, because the production function is sufficiently flexible in log-skills. In addition, dividing by the standard deviation, implies that the objective is invariant to the scale.

3.2.4 Anchoring

Consider equations (8)–(12) and recall that $\tilde{\mu}_{\theta,t,0} = \tilde{\mu}_{I,t,0} = 0$ and $\tilde{\lambda}_{\theta,t,0} = \tilde{\lambda}_{I,t,0} = 1$, and the parameters in this system of equations are point identified by Theorem 1. Now define $\tilde{\vartheta}_t$ such that $\ln \tilde{\vartheta}_t = \tilde{\rho}_0 + \tilde{\rho}_1 \ln \tilde{\theta}_t$. We then get

$$\begin{aligned}
 \ln \tilde{\vartheta}_{t+1} &= \tilde{\rho}_0 + \tilde{\rho}_1 \tilde{a}_t - \tilde{\rho}_0 \tilde{\gamma}_{1t} + \tilde{\gamma}_{1t} \ln \tilde{\vartheta}_t + \tilde{\rho}_1 \tilde{\gamma}_{2t} \ln I_t + \tilde{\gamma}_{3t} \ln \tilde{\vartheta}_t \ln I_t + \tilde{\rho}_1 \tilde{\eta}_{\theta,t} \\
 Z_{\theta,t,m} &= \tilde{\mu}_{\theta,t,m} - \frac{\tilde{\rho}_0 \tilde{\lambda}_{\theta,t,m}}{\tilde{\rho}_1} + \frac{\tilde{\lambda}_{\theta,t,m}}{\tilde{\rho}_1} \ln \tilde{\vartheta}_t + \varepsilon_t \\
 Z_{I,t,m} &= \tilde{\mu}_{I,t,m} + \tilde{\lambda}_{I,t,m} \ln \tilde{I}_t + \varepsilon_{I,t,m} \\
 \ln \tilde{I}_t &= \tilde{\beta}_{0t} - \frac{\tilde{\rho}_0 \tilde{\beta}_{1t}}{\tilde{\rho}_1} + \frac{\tilde{\beta}_{1t}}{\tilde{\rho}_1} \ln \tilde{\vartheta}_t + \tilde{\beta}_{2t} \ln Y_t + \tilde{\eta}_{I,t} \\
 Q &= \ln \tilde{\vartheta}_T + \eta_Q
 \end{aligned}$$

Cunha and Heckman (2008) estimate a model for $\ln \tilde{\vartheta}_t$ which anchors the skills at Q . Their identification strategy is equivalent to using Assumptions 2 and 3(a) and imposing $\rho_0 = 0$ and $\rho_1 = 1$ instead of Assumption 2. Anchoring can help with the interpretation of certain parameters of the model. For example, when $E(Q \mid \ln \tilde{\vartheta}_T) = \ln \tilde{\vartheta}_T$, then an increase of $\ln \tilde{\vartheta}_T$ by one corresponds to a one unit increase in the conditional expectation of Q . However, as illustrated in Example 2, investment or income sequences that maximize the expected levels of skills depend on the units of measurements of the anchor.⁵

As discussed by Cunha and Heckman (2008), $\tilde{\gamma}_{1t}$ is invariant to the anchor. However, without Assumption 3(a), $\tilde{\gamma}_{1t}$ is typically not equal to γ_{1t} . The coefficient in front of investment is hard to interpreted because it depends on the specific anchor and its units of measurements (as also noted by Cunha and Heckman (2008)). Moreover, skills can be anchored at the adult outcome, but not investment, and many of the estimated parameters also depend on the units of the investment measures. While anchoring makes most sense under Assumption 3(a), in which case the skills in all periods are in the units of Q , we could instead use Assumption 3(b) to achieve point identification. In this case different adult outcomes or different units of measurements lead to different coefficients that possibly suggest very different dynamics, just as in Example 1.

3.3 CES production function

I now discuss the CES production technology, where

$$(13) \quad \theta_{t+1} = (\gamma_{1t}\theta_t^{\sigma_t} + \gamma_{2t}I_t^{\sigma_t})^{\frac{\psi_t}{\sigma_t}} \exp(\eta_{\theta,t}) \quad t = 0, \dots, T-1$$

$$(14) \quad Z_{\theta,t,m} = \mu_{\theta,t,m} + \lambda_{\theta,t,m} \ln \theta_t + \varepsilon_{\theta,t,m} \quad t = 0, \dots, T, m = 1, 2$$

$$(15) \quad Z_{I,t,m} = \mu_{I,t,m} + \lambda_{I,t,m} \ln I_t + \varepsilon_{I,t,m} \quad t = 0, \dots, T-1, m = 1, 2$$

$$(16) \quad \ln I_t = \beta_{0t} + \beta_{1t} \ln \theta_t + \beta_{2t} \ln Y_t + \eta_{I,t} \quad t = 0, \dots, T-1$$

$$(17) \quad Q = \rho_0 + \rho_1 \ln \theta_T + \eta_Q$$

⁵The identification arguments are fundamentally different if the anchoring equations was in levels instead of logs of the skills. That is, if $Q = \rho_0 + \rho_1 \theta_T + \eta_Q = \rho_0 + \frac{\rho_1}{\exp(\mu_{\theta,T,1})^{1/\lambda_{\theta,T,1}}} \tilde{\theta}_T^{1/\lambda_{\theta,T,1}} + \eta_Q$. In this case it can be shown that the joint distribution of $(Q, \tilde{\theta}_T)$ is identified, which identifies $\lambda_{\theta,T,1}$. Thus, the distribution of θ_T is identified up to a scaling factors, which implies that even sequences of income or investment that maximize the level of skills are identified.

where $\sigma_t \neq 0$, $\gamma_{1t} \neq 0$, and $\gamma_{2t} \neq 0$ for all t . The measurement system is linear in $\ln \theta_t$ (as in Cunha et al. (2010) or Attanasio et al. (2020)) to ensure that estimated skills are positive.

3.3.1 Identification

Similar to the trans-log case, define $\tilde{\theta}_t = \exp(\mu_{\theta,t,1})\theta_t^{\lambda_{\theta,t,1}}$ so that

$$\theta_t = \exp\left(-\frac{\mu_{\theta,t,1}}{\lambda_{\theta,t,1}}\right)\tilde{\theta}_t^{\frac{1}{\lambda_{\theta,t,1}}}.$$

Similarly, let $\tilde{I}_t = \exp(\mu_{I,t,1})I_t^{\lambda_{I,t,1}}$. We can again rewrite the production technology in terms of $\tilde{\theta}_t$ and \tilde{I}_t . That is, we can rewrite equations (13)–(17) to

$$(18) \quad \tilde{\theta}_{t+1} = \left(\tilde{\gamma}_{1t}\tilde{\theta}_t^{\frac{\sigma_t}{\lambda_{\theta,t,1}}} + \tilde{\gamma}_{2t}\tilde{I}_t^{\frac{\sigma_t}{\lambda_{I,t,1}}}\right)^{\frac{\lambda_{\theta,t+1,1}\psi_t}{\sigma_t}} \exp(\tilde{\eta}_{\theta,t}) \quad t = 0, \dots, T-1$$

$$(19) \quad Z_{\theta,t,m} = \tilde{\mu}_{\theta,t,m} + \tilde{\lambda}_{\theta,t,m} \ln \tilde{\theta}_t + \varepsilon_{\theta,t,m} \quad t = 0, \dots, T, m = 1, 2$$

$$(20) \quad Z_{I,t,m} = \tilde{\mu}_{I,t,m} + \tilde{\lambda}_{I,t,m} \ln \tilde{I}_t + \varepsilon_{I,t,m} \quad t = 0, \dots, T-1, m = 1, 2$$

$$(21) \quad \ln \tilde{I}_t = \tilde{\beta}_{0t} + \tilde{\beta}_{1t} \ln \tilde{\theta}_t + \tilde{\beta}_{2t} \ln Y_t + \tilde{\eta}_{I,t} \quad t = 0, \dots, T-1$$

$$(22) \quad Q = \tilde{\rho}_0 + \tilde{\rho}_1 \ln \tilde{\theta}_T + \eta_Q$$

where

$$\tilde{\gamma}_{1t} = \gamma_{1t} \exp\left(\sigma_t \left(\frac{\mu_{\theta,t+1,1}}{\lambda_{\theta,t+1,1}\psi_t} - \frac{\mu_{\theta,t,1}}{\lambda_{\theta,t,1}}\right)\right) \quad \text{and} \quad \tilde{\gamma}_{2t} = \gamma_{2t} \exp\left(\sigma_t \left(\frac{\mu_{\theta,t+1,1}}{\lambda_{\theta,t+1,1}\psi_t} - \frac{\mu_{I,t,1}}{\lambda_{I,t,1}}\right)\right)$$

and the other parameters are defined as in the trans-log case. Using the identified joint distribution of $(\{\ln \tilde{\theta}_t\}_{t=0}^T, \{\ln \tilde{I}_t\}_{t=0}^T)$ and the restrictions imposed by the production function, the following theorem characterizes the identified set of the finite dimensional parameters.

Theorem 3. *Suppose Assumption 1 holds.*

- (a) *The identified set of $\{\mu_{\theta,t,m}, \lambda_{\theta,t,m}\}_{t=0,\dots,T,m=1,2}$, $\{\mu_{I,t,m}, \lambda_{I,t,m}\}_{t=0,\dots,T-1,m=1,2}$, (ρ_0, ρ_1) , $\{\beta_{0t}, \beta_{1t}, \beta_{2t}\}_{t=0}^{T-1}$, $\{\gamma_{1t}, \gamma_{2t}, \sigma_t, \psi_t\}_{t=0}^{T-1}$ consists of all vectors that yield the same values of $\{\tilde{\mu}_{\theta,t,m}, \tilde{\lambda}_{\theta,t,m}\}_{t=0,\dots,T,m=1,2}$, $\{\tilde{\mu}_{I,t,m}, \tilde{\lambda}_{I,t,m}\}_{t=0,\dots,T-1,m=1,2}$, $(\tilde{\rho}_0, \tilde{\rho}_1)$, $\{\tilde{\gamma}_{1t}, \tilde{\gamma}_{2t}, \frac{\sigma_t}{\lambda_{\theta,t,1}}, \frac{\sigma_t}{\lambda_{I,t,1}}\}_{t=0}^{T-1}$, $\{\frac{\sigma_t\psi_t}{\lambda_{\theta,t+1,1}}\}_{t=0}^{T-1}$, and $\{\tilde{\beta}_{0t}, \tilde{\beta}_{1t}, \tilde{\beta}_{2t}\}_{t=0}^{T-1}$ as the true parameter vectors.*

- (b) Let $\{\bar{\mu}_{\theta,t,1}\}_{t=0}^T$, $\{\bar{\mu}_{I,t,1}\}_{t=0}^{T-1}$, and $\{\bar{\lambda}_{\theta,t,1}\}_{t=0}^T$ be fixed constants with $\bar{\lambda}_{\theta,t,1} \neq 0$ for all t . Under the additional restrictions $\{\mu_{\theta,t,1}\}_{t=0}^T = \{\bar{\mu}_{\theta,t,1}\}_{t=0}^T$ and $\{\mu_{I,t,1}\}_{t=0}^{T-1} = \{\bar{\mu}_{I,t,1}\}_{t=0}^{T-1}$ and $\{\lambda_{\theta,t,1}\}_{t=0}^T = \{\bar{\lambda}_{\theta,t,1}\}_{t=0}^T$ the identified set is a singleton.

An important implications of part (a) is that the fraction $\frac{\lambda_{\theta,t,1}}{\lambda_{I,t,1}} = \frac{\lambda_{\theta,t,1}}{\sigma_t} \frac{\sigma_t}{\lambda_{I,t,1}}$ is point identified for all $t = 1, \dots, T-1$. Hence, if we restrict $\lambda_{\theta,t,1}$ to a constant, $\lambda_{I,t,1}$ is identified, which is very different to the trans-log case. Intuitively, since skills and investment have the same exponent in the CES production functions, the relative scale is identified by the functional form restrictions.

As before, I now introduce additional assumptions to achieve point identification.

Assumption 2'. $\lambda_{\theta,0,1} = 1$ and $\mu_{\theta,0,1} = 0$.

Assumption 3'.

- (a) $\mu_{\theta,t,1} = \mu_{\theta,t+1,1}$ for all $t = 0, \dots, T-1$
- (b) $\gamma_{1t} + \gamma_{2t} = 1$ for all $t = 0, \dots, T-1$.

Assumption 4'.

- (a) $\mu_{I,t,1} = \mu_{I,t+1,1} = 0$ for all $t = 0, \dots, T-2$.
- (b) $\beta_{0t} = 0$ for all $t = 0, \dots, T-1$.

Assumption 5'.

- (a) $\lambda_{\theta,t,1} = \lambda_{\theta,t+1,1}$ for all $t = 0, \dots, T-1$.
- (b) $\psi_t = 1$ for all $t = 0, \dots, T-1$.

The following result shows how point identification can be established.

Corollary 2. *Suppose Assumptions 1 and 2' hold. Suppose either Assumption 3'(a) or Assumption 3'(b) holds. Suppose either Assumption 4'(a) or Assumption 4'(b) holds. Suppose either Assumption 5'(a) or Assumption 5'(b) holds. Then all parameters are point identified.*

A common restriction with the CES production function is to set $\psi_t = 1$ for all t (see e.g. Cunha et al. (2010) and Attanasio et al. (2020)). In this case, Theorem 3 implies that $\frac{\lambda_{\theta,t+1,1}}{\lambda_{\theta,t,1}}$ is point identified. Hence, assuming age-invariance (i.e. $\lambda_{\theta,t+1,1} = \lambda_{\theta,t,1}$) is not required, and it is in fact testable. Moreover, with $\psi_t = 1$, Corollary 2 implies that the only scale restriction needed is $\lambda_{\theta,0,1} = 1$ (or alternatively $\lambda_{I,0,1} = 1$). Nevertheless, it is common practice to set $\lambda_{\theta,t,1} = \lambda_{I,t,1} = 1$ for all t , which are not normalizations, but restrictive assumptions (even if $\psi_t \neq 1$). Setting the scales to different numbers or changing the units of measurement of the data then affects all estimated parameters, optimal investment sequences, and other counterfactuals. Similarly, if the scale of investment is fixed, we cannot anchor the skills at the adult outcome, unless $\rho_1 = 1$, which can only be true for one specific unit of measurements (see Appendix B.1). The exact consequences of imposing unnecessary restrictions depend on the estimation methods used, and I discuss specific examples in Sections 4 and 5.

As shown in Theorem A.2 in the appendix and illustrated in examples below, if the model is correctly specified and Assumption 1 holds, then there always exist sets of parameters which are consistent with the data and satisfy Assumptions 1, 2', either 3'(a) or 3'(b), either 4'(a) or 4'(b), and either 5'(a) or 5'(b). Similar to the trans-log case, different sets of assumptions yield observationally equivalent models with potentially very different primitive parameters.

3.3.2 Invariant parameters

As in the trans-log case, important policy relevant parameters are point identified under Assumption 1 only, because they can all be written in terms of identified features. Similar to before, now define $s_{1t}(\alpha_1, \alpha_2, \alpha_3) = (\gamma_{1t}Q_{\alpha_1}(\theta_t)^{\sigma_t} + \gamma_{2t}Q_{\alpha_2}(I_t)^{\sigma_t})^{\frac{\psi_t}{\sigma_t}} \exp(Q_{\alpha_3}(\eta_{\theta,t}))$ and $s_{2t}(\alpha_1, \alpha_2, \alpha_3, y) = (\gamma_{1t}Q_{\alpha_1}(\theta_t)^{\sigma_t} + \gamma_{2t}I_t(y)^{\sigma_t})^{\frac{\psi_t}{\sigma_t}} \exp(Q_{\alpha_2}(\eta_{\theta,t}))$ where $\ln I_t(y) = \beta_{0t} + \beta_{1t} \ln Q_{\alpha_1}(\theta_t) + \beta_{2t} \ln y + Q_{\alpha_3}(\eta_{I,t})$.

Theorem 4. *Suppose Assumption 1 holds.*

1. $F_{\theta_{t+1}}(s_{1t}(\alpha_1, \alpha_2, \alpha_3))$ and $\mu_{\theta,t+1,m} + \lambda_{\theta,t+1,m} \ln s_{1t}(\alpha_1, \alpha_2, \alpha_3) + Q_{\alpha_4}(\varepsilon_{\theta,t+1,m})$ are point identified for all $\{\alpha_j\}_{j=1}^4 \in (0, 1)$. Moreover, $\frac{\partial \lambda_{\theta,t+1,m} \ln \theta_{t+1}}{\partial \lambda_{\theta,t,m'} \ln \theta_t} \big|_{\theta_t=Q_{\alpha_1}(\theta_t), I_t=Q_{\alpha_2}(I_t)}$ and $\frac{\partial \lambda_{\theta,t+1,m} \ln \theta_{t+1}}{\partial \lambda_{I,t,m'} \ln I_t} \big|_{\theta_t=Q_{\alpha_1}(\theta_t), I_t=Q_{\alpha_2}(I_t)}$ are point identified for all m, m' and $\{\alpha_j\}_{j=1}^2 \in (0, 1)$.

2. $F_{\theta_{t+1}}(s_{2t}(\alpha_1, \alpha_2, \alpha_3, y))$ and $\mu_{\theta,t+1,m} + \lambda_{\theta,t+1,m} \ln s_{2t}(\alpha_1, \alpha_2, \alpha_3, y) + Q_{\alpha_4}(\varepsilon_{\theta,t+1,m})$ are point identified $\{\alpha_j\}_{j=1}^4 \in (0, 1)$.
3. $P(Q \leq q \mid \theta_t = Q_{\alpha_1}(\theta_t), \{I_s = Q_{\alpha_{2s}}(I_s)\}_{s=t}^{T-1}, \{\eta_{\theta,s} = Q_{\alpha_{3s}}(\eta_{\theta,s})\}_{s=t}^{T-1})$ is point identified for all $\alpha_1, \{\alpha_{2s}, \alpha_{3s}\}_{s=t}^{T-1} \in (0, 1)$.
4. $P(Q \leq q \mid \theta_0 = Q_{\alpha}(\theta_0), \{Y_t = y_t\}_{t=0}^{T-1})$ is point identified for all $\alpha \in (0, 1)$.
5. If in addition either Assumption 5'(a) or Assumption 5'(b) holds, the distributions of

$$\frac{\partial \ln \theta_{t+1}}{\partial \ln \theta_t} = \frac{\partial}{\partial \ln \theta_t} \ln (\gamma_{1t} \theta_t^{\sigma_t} + \gamma_{2t} I_t^{\sigma_t})^{\frac{1}{\sigma_t}} \quad \text{and} \quad \frac{\partial \ln \theta_{t+1}}{\partial \ln I_t} = \frac{\partial}{\partial \ln I_t} \ln (\gamma_{1t} \theta_t^{\sigma_t} + \gamma_{2t} I_t^{\sigma_t})^{\frac{1}{\sigma_t}}$$

are point identified and

$$\left. \frac{\partial \ln \theta_{t+1}}{\partial \ln \theta_t} \right|_{\ln \theta_t = Q_{\alpha_1}(\ln \theta_t), \ln I_t = Q_{\alpha_2}(\ln I_t)} = \left. \frac{\partial}{\partial \ln \theta_t} \ln (\gamma_{1t} \theta_t^{\sigma_t} + \gamma_{2t} I_t^{\sigma_t})^{\frac{1}{\sigma_t}} \right|_{\ln \theta_t = Q_{\alpha_1}(\ln \theta_t), \ln I_t = Q_{\alpha_2}(\ln I_t)}$$

and

$$\left. \frac{\partial \ln \theta_{t+1}}{\partial \ln I_t} \right|_{\ln \theta_t = Q_{\alpha_1}(\ln \theta_t), \ln I_t = Q_{\alpha_2}(\ln I_t)} = \left. \frac{\partial}{\partial \ln I_t} \ln (\gamma_{1t} \theta_t^{\sigma_t} + \gamma_{2t} I_t^{\sigma_t})^{\frac{1}{\sigma_t}} \right|_{\ln \theta_t = Q_{\alpha_1}(\ln \theta_t), \ln I_t = Q_{\alpha_2}(\ln I_t)}$$

are point identified for all $\alpha_1, \alpha_2 \in (0, 1)$.

The features in parts (1)–(4) are analogous to those in the trans-log case and can be used to calculate optimal investment/income strategies and anchored treatment effects, as discussed after Theorem 2. As opposed to the trans-log case, now the relative scales of skills and investment are point identified. Consequently, elasticities are identified under Assumption 1 and either age-invariant skill measures or $\psi_t = 1$ only.

3.3.3 Non-invariant parameters

To achieve point identification of all parameters, we need to fix the levels of the logs of skills and investment, e.g. by setting $\mu_{\theta,0,1}$ and $\mu_{0,I,1}$ to 0. Similar to the trans-log case, these restrictions are generally not normalizations for the primitive parameters. In the example below, I illustrate that with the restriction $\gamma_{1t} + \gamma_{2t} = 1$, setting $\mu_{\theta,0,1} = 0$ can imply different

dynamics. Here $\mu_{\theta,0,1}$ (and not $\lambda_{\theta,0,1}$) affects the scale because we can identify the distribution of $\tilde{\theta}_t = \exp(\mu_{\theta,t,1})\theta_t^{\lambda_{\theta,t,1}}$ and the production function is in levels rather than logs of θ_t .

Example 3. The issues with the restriction $\mu_{\theta,0,1} = 0$ in the CES case are analogous to the issues with the restriction $\lambda_{\theta,0,1} = 1$ in the trans-log case discussed in Example 1. I now consider a numerical example analogous to Example 1 and focus on the measurement and the production function only. That is, suppose that

$$\begin{aligned}\theta_{t+1} &= \left(\frac{1}{2}\theta_t + \frac{1}{2}I_t\right) \exp(\eta_{\theta,t}) \\ Z_{\theta,t,m} &= \ln(12) + \ln \theta_t + \varepsilon_{\theta,t,m}\end{aligned}$$

Notice that $Z_{\theta,t,m} = \ln(12\theta_t) + \varepsilon_{\theta,t,m}$. Let $\tilde{\theta}_t = 12\theta_t$. Just as in Example 1, we get

$$\begin{aligned}\tilde{\theta}_{t+1} &= \left(\tilde{\gamma}_{1t}\tilde{\theta}_t + (1 - \tilde{\gamma}_{1t})I_t\right) \exp(\tilde{\eta}_{\theta,t}) \\ Z_{\theta,t,m} &= \tilde{\mu}_{\theta,t,m} + \ln \tilde{\theta}_t + \varepsilon_{\theta,t,m}\end{aligned}$$

where $(\tilde{\gamma}_{10}, \tilde{\gamma}_{11}, \tilde{\gamma}_{12}, \tilde{\gamma}_{13}, \dots) = (0.077, 0.351, 0.435, 0.470, \dots)$ and

$$(\exp(\tilde{\mu}_{0,\theta,m}), \exp(\tilde{\mu}_{1,\theta,m}), \exp(\tilde{\mu}_{2,\theta,m}), \exp(\tilde{\mu}_{3,\theta,m}), \dots) = (1, 6.5, 9.25, 10.625, \dots).$$

Just as in Example 1, these two models are observationally equivalent, but suggest very different dynamics.

4 Monte Carlo simulations

I use a very similar data generating process as Attanasio et al. (2020). In particular, I use

$$\theta_{t+1} = A_t (\gamma_t \theta_t^{\sigma_t} + (1 - \gamma_t) I_t^{\sigma_t})^{\frac{1}{\sigma_t}} \exp(\eta_{\theta,t})$$

for $t = 1, 2$. Allowing for $A_t \neq 1$ is equivalent to not imposing the restriction that the coefficients in front of θ_t and I_t sum to 1. Since I want to study income transfers as counterfactuals, I augment the setup of Attanasio et al. (2020) and add $\ln I_t = \beta_{1t} \ln \theta_t + \beta_{2t} \ln Y_t + \eta_{I,t}$, where

$Y_0 = Y_1$. To simulate data, I first draw $(\ln(\theta_0), \ln(Y))$ from a normal mixture distribution. Given $(\ln(\theta_0), \ln(Y))$ and normally distributed $\eta_{\theta,t}$ and $\eta_{I,t}$ I then generate I_0 , θ_1 , I_1 , and θ_2 using the model. If $\ln I_t$ was equal to $\ln Y_t$, the setup would be exactly equal to that of Attanasio et al. (2020) with parameters as in their Table 9, which are based on their empirical results, and I use $\sigma_0 = \sigma_1 = -0.5$.⁶ I deviate slightly from their setting by using the additional investment equation with $\beta_{1t} = 0.1$, $\beta_{2t} = 0.9$, and $\eta_{I,t} \sim N(0, 0.1^2)$. I simulate three measures each for θ_t and I_t , which have a factor structure with $\mu_{\theta,t,m} = \mu_{I,t,m} = 0$ for all m, t . In addition $\lambda_{\theta,t,1} = \lambda_{I,t,1} = 1$ for all t , which imposes the scale restrictions and the age-invariance assumptions. In Attanasio et al. (2020), all of these loadings are also “normalized” to 1 in their estimation procedure. In particular, they first estimate the distribution of the measures and of log-income using a normal mixture model. Then, assuming that $(\ln(\theta_0), \ln(\theta_1), \ln(\theta_2), \ln(I_0), \ln(I_1), \ln(Y))$ also has a normal mixture distribution, they use the factor structure and the restrictions to estimate that distribution.⁷ Lastly, they take draws from the distribution and estimate the production function parameters by nonlinear least squares. I implement their estimator and my more flexible approach.

For my approach, I use the estimation procedure explained in Appendix C.2, which simplifies in this setup because investment is exogenous (and thus, $\kappa_t = 0$). Moreover, to focus on the scale restriction only, I set $\mu_{\theta,t,m} = \mu_{I,t,m} = 0$ for all m and t . Finally, I impose that the first skill measure and the first investment measure are age-invariant. I then set $\lambda_{I,t,1} = 1$ for all t , and estimate $\lambda_{\theta,t,1}$. I therefore impose Assumptions 2', 3'(a), and 4'(a), as well as both parts of Assumption 5'. While only one part of the last assumption is needed (i.e. age-invariance of the measures could be dropped), they are both satisfied in this setup. I then estimate $\lambda_{\theta} = \lambda_{\theta,t,1}$ along with the production function parameters by solving

$$\arg \min_{\lambda_{\theta,1}, \sigma_1, \sigma_2, \gamma_{11}, \gamma_{21}, \gamma_{11}, \gamma_{21}} \sum_{t=1}^T \sum_{j=1}^J \left(\ln \tilde{\theta}_{t+1,j} - \ln(A_t) \lambda_{\theta} - \frac{\lambda_{\theta}}{\sigma_t} \ln \left(\gamma_{1t} \tilde{\theta}_{t,j}^{\frac{\sigma_t}{\lambda_{\theta}}} + (1 - \gamma_{2t}) \tilde{I}_{t,j}^{\sigma_t} \right) \right)^2$$

⁶I use slightly different notation to be consistent with the notation above. Specifically, the periods are $t = 0, 1, 2$ instead of $t = 1, 2, 3$, I use I_t instead of X_t for the second latent variable, and I use σ_t instead of ρ_t to denote the elasticity of substitution. The distribution of $(\ln(\theta_0), \ln(Y))$ is the same as the distribution of $(\ln(\theta_0), \ln(X))$ in Attanasio et al. (2020).

⁷Interestingly, one can see from the DGP that $(\ln(\theta_0), \ln(\theta_1), \ln(\theta_2))$ does not have a normal mixture distribution due to the nonlinear production function, but this misspecification bias seems to be relatively small in this simulation setup.

where $\tilde{\theta}_{t,j}$ and $\tilde{I}_{t,j}$ are draws from the estimated distribution of skills and investment using the estimator of Attanasio et al. (2020). Similarly, we can estimate β_{1t} and β_{2t} from a regression of $\tilde{I}_{t,j}$ on $\tilde{\theta}_{t,j}$ and Y_j .

In the following, I will investigate the effect of multiplying the skill measures by a single constant s_θ in all periods, which changes the loadings, but not the age-invariance assumption. For example, the first measure, say $\tilde{Z}_{\theta,t,1}$, is generated by

$$\tilde{Z}_{\theta,t,1} = \log(\theta_t) + \varepsilon_{\theta,t,1}$$

but when I estimate the model, I use

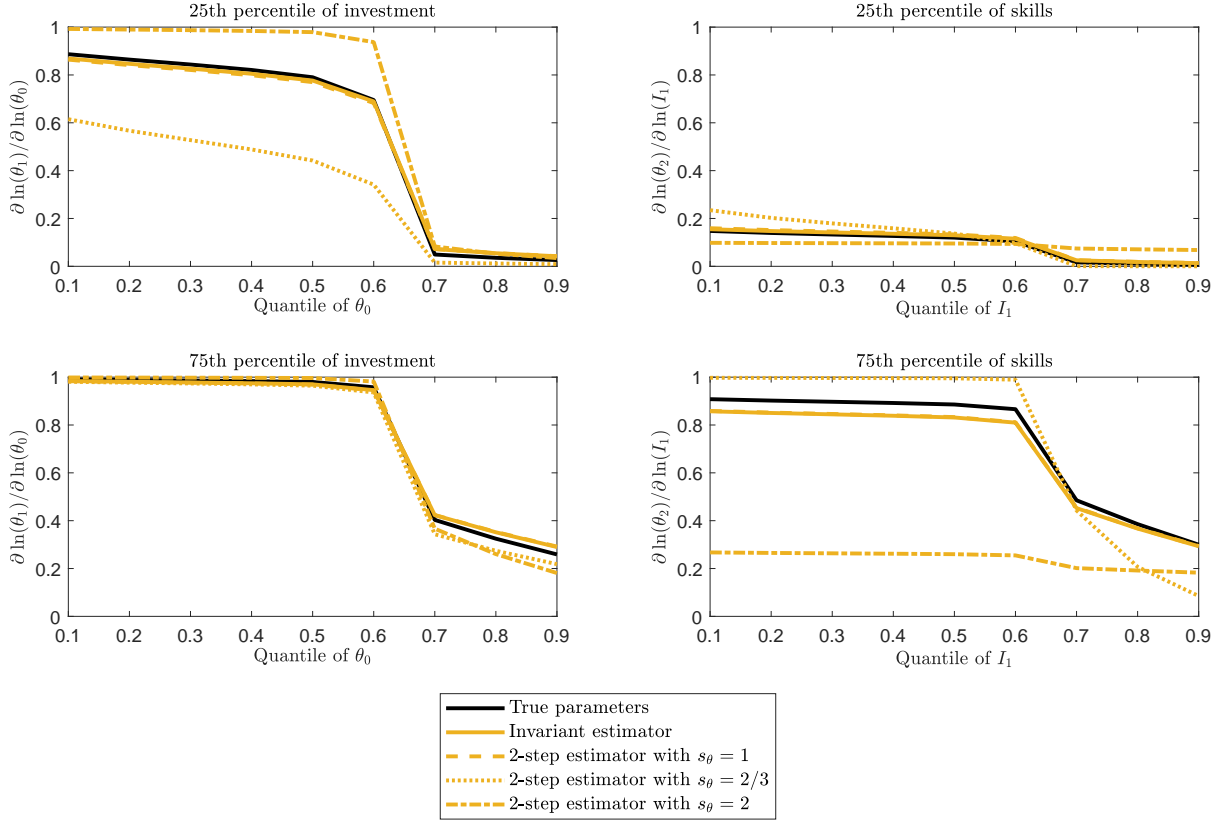
$$Z_{\theta,t,1} = s_\theta \tilde{Z}_{\theta,t,1} = s_\theta \log(\theta_t) + s_\theta \varepsilon_{\theta,t,1},$$

which is also an age-invariant measure. The estimators still impose that the loadings are equal to 1 ($\lambda_{I,t,1} = 1$ for all t with my estimator and $\lambda_{\theta,t,1} = 1$ and $\lambda_{I,t,1} = 1$ for all t with the estimator of Attanasio et al. (2020)). Of course, in practice, we do not know the DGP and there is no reason to believe that the true loading is 1. Ideally, the restriction should be a normalization in which case the results would be invariant to scaling the data or changing its units of measurement. However, Corollary 2 implies that the estimator of Attanasio et al. (2020) is inconsistent. I consider the implications for elasticities and counterfactual predictions, which are point identified (as shown in Theorem 4) and are invariant to the scaling when using the new estimator.⁸ I take $s_\theta = 1$, which is the correctly specified model, as well as $s_\theta = 2/3$ and $s_\theta = 2$. I report average estimates from 500 simulated data sets.

To summarize the production function estimates I report $\partial \ln(\theta_1)/\partial \ln(\theta_0)$ as a function of quantiles of $\ln(\theta_0)$ and evaluated at the 25th and 75th percentile of I_0 as well as $\partial \ln(\theta_2)/\partial \ln(I_1)$ as a function of quantiles of $\ln(I_1)$ and evaluated at the 25th and 75th percentile of θ_1 . Figure 2 displays these partial derivatives for the true parameters, the 2-step estimator of Attanasio et al. (2020) with different values of s_θ , and the invariant estimator that also estimates the scale. In this and the following figures, the results obtained with the

⁸These features are also invariant to scaling the investment measures with my estimator and I could have set $\lambda_{\theta,t,1}$ instead of $\lambda_{I,t,1}$ to 1.

Figure 2: Partial derivatives

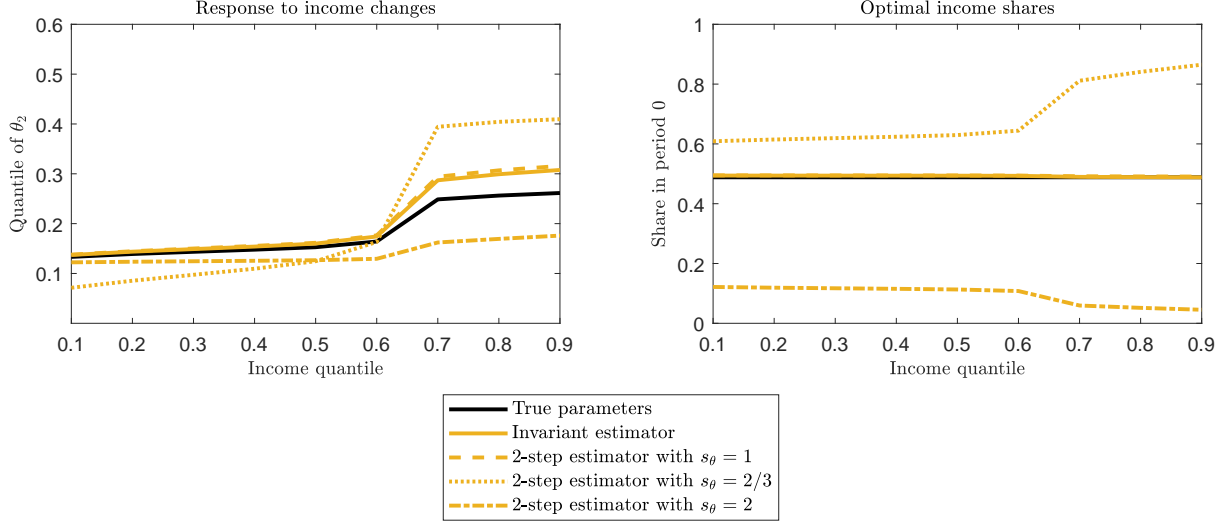


Notes: This figure shows $\partial \ln(\theta_1) / \partial \ln(\theta_0)$ as a function of quantiles of $\ln(\theta_0)$ and evaluated at the 25th and 75th percentile of I_0 as well as $\partial \ln(\theta_2) / \partial \ln(I_1)$ as a function of quantiles of $\ln(I_1)$ and evaluated at the 25th and 75th percentile of θ_1 using different estimators.

invariant estimator are always almost identical to those with the 2-step estimator and $s_\theta = 1$ and very similar to those with the true parameter values. Moreover, the invariant estimator adapts to the scale change. Contrarily, the figure illustrates that small changes in the units of measurements can have a large effect on the results when using the 2-step estimator. One interesting finding is that the 2-step estimator underestimates the partial derivative with respect to $\ln(\theta_0)$ at the 25th percentile of I_0 for $s_\theta = 2/3$ and overestimates it for $s_\theta = 2$.

As a first counterfactual, consider an individual, whose value of θ_0 equals $Q_{0.1}(\theta_0)$ and all unobservables are 0. I then exogenously change the income sequence of that individual and check the implied quantile in the skill distribution in period 2. Of course, the larger income, the higher the relative rank/quantile in the last period. I report results where a feasible choice of income in each period is a given quantile (that will be varied). Instead of using the

Figure 3: Response to income changes and optimal shares



Notes: The figure shows the estimated effects of increases in income on the quantiles of skills in the last periods using different estimators. The left panel shows the quantiles for different income levels and the right panels shows the corresponding optimal income shares in period 0.

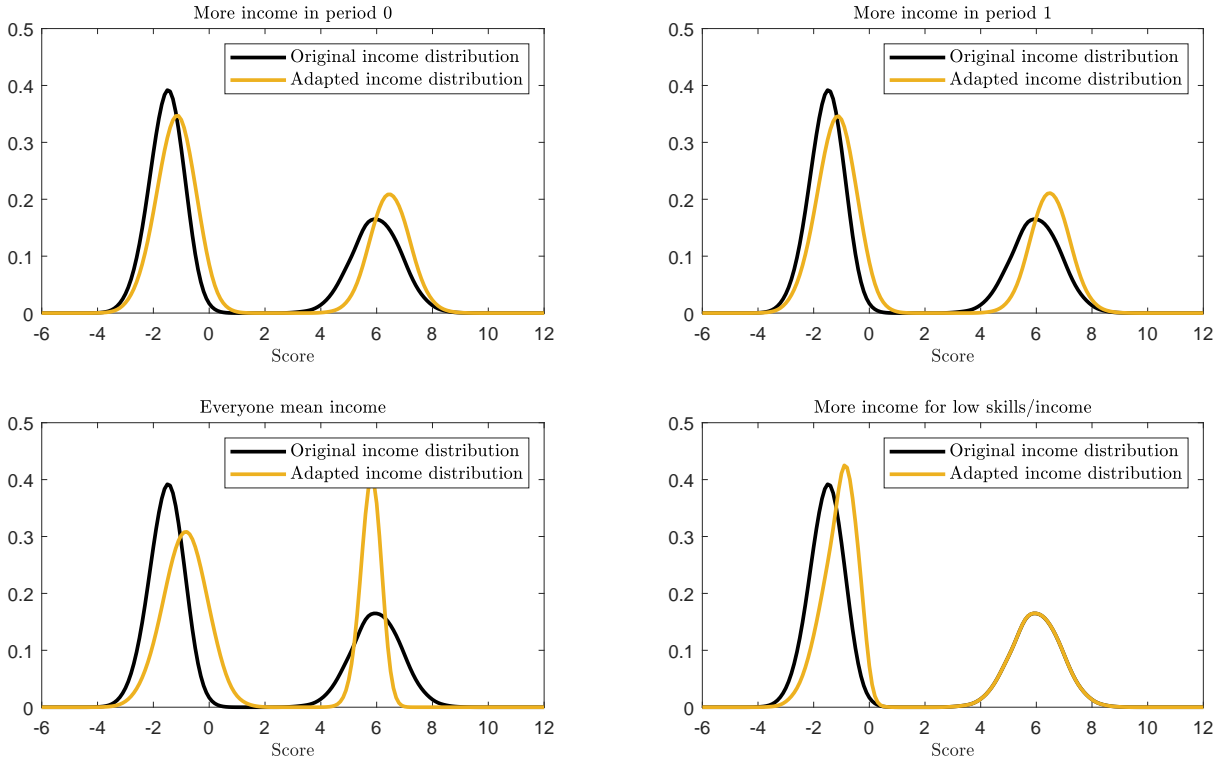
feasible choices, I distribute the total income among the two periods to maximize the skill quantile in the last period. The second part of Theorem 4 implies that these counterfactuals are point identified, can be consistently estimated with the invariant estimator, but the results with the 2-step estimator of Attanasio et al. (2020) will depend on s_θ .

The left panel of Figure 3 shows the quantile as a function of the feasible income quantile (0.5 is the median, etc.). Using the true parameters, we can see that, even for large income, the quantile in period 2 is at most around 0.26 and is almost flat below the median. The right panel of Figure 3 displays the corresponding optimal income shares in period 0. With the true parameters, income should be similar in both periods. The 2-step estimator with $s_\theta = 1$ and the invariant estimator (irrespective of the scale) yield similar conclusions for the optimal income sequence, but have a small bias for the estimated quantile in period 2 (due the approximation error of the joint distribution of the measures/skills as mixtures of normals). Figure 3 also demonstrates that changing the units of measurement can lead to inefficient income transfers when using the 2-step estimator. For example, with $s_\theta = 2/3$ the results suggest that we should mainly invest in $t = 0$, and with $s_\theta = 2$ it appears that we should mainly invest in $t = 1$. Moreover, the estimated gains of income transfers are

misleading in this case. The inconsistent estimates in the left panel suggest that large income transfers can increase the quantile to almost 0.5 in period 2 when $s_\theta = 2/3$. Contrarily, with $s_\theta = 2$ we would underestimate the effect. Importantly, the results for the new estimator are invariant to changes in the units of measurements.

Next, I consider how exogenous income changes affect the skill distribution. To do so, I take draws from the estimated joint distribution of income and skills in period 0 (based on the average of the estimated parameters to get representative results) and consider four counterfactual marginal income distributions. First, I increase everyone's income by two standard deviations in period 0. Second, I increase everyone's income by two standard deviations in period 1. Third, I set income to the median for everyone in both periods. Fourth, I increase income by two standard deviations in both periods, but only if the initial skill and income quantiles are below 0.5. I set all unobservables to their median values. I

Figure 4: Outcome distributions under different investment sequences



Notes: The figure shows counterfactual outcome distributions. Each panel displays the distribution with the original income distribution and a counterfactual income distribution.

report results for the invariant estimator only.

Figure 4 shows the implied distributions of one of the skill measures in the final period (which could be a test score or an adult outcome in an application). These results depend on and should be interpreted relative to the units of measurements of that measures. Figure 4 is based on $s_\theta = 1$. One can see that increasing income in either of the first two periods has very similar effects and leads to an increase in skills. Since everyone’s income increases, everyone is better off. If everyone receives the mean income, the variance of the outcome distribution decreases considerably. If we only increase income for people with low initial skills and income, then predictably only the lower tail of the distribution will be affected.

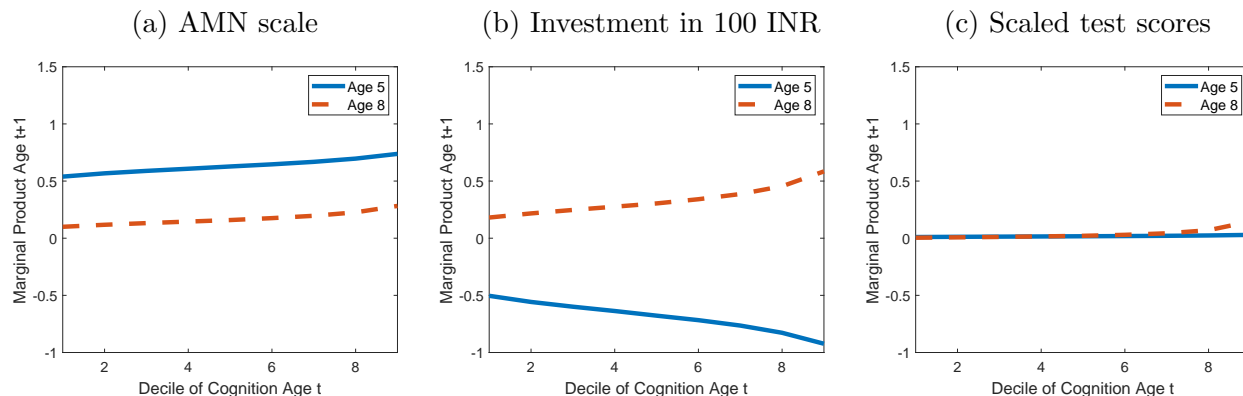
5 Empirical illustration

In this section, I illustrate the previous findings by replicating some of the estimates of Attanasio et al. (2020) and showing how changing the units of measurements of the data (or, equivalently, changing the scale restriction) affects parameter estimates and counterfactuals.

Attanasio et al. (2020) estimate production functions for cognitive skills and health of children with past skills and health as well as parental investment, health, and skills as inputs. They use the CES production function with $\psi_t = 1$. Theorem 3 and Corollary 2 imply that most scale parameters are identified and one only has to restrict the loading of one latent variable in one time period to 1. Attanasio et al. (2020) set a loading for each observed factor in each time period to 1. Specifically, they write on page 2520: “One way to meet this condition is to normalize each factor on the same measure every period, assuming that the mapping from measure to factor is invariant with respect to the age of the subject. Fortunately, our data are sufficiently rich that we are able to do this for our model. For child cognitive skills, we always normalize the loading on PPVT to one. Similarly, child health is always normalized on height z scores, investments are normalized on amount spent on books, parental health is normalized on mother’s weight, parental cognitive skills is normalized on mother’s years of schooling, and resources are normalized on income.” These restrictions already take the considerations of Agostinelli and Wiswall (2016a, 2024) into account.

I consider two changes of the units of measurements. First, I scale book expenditures

Figure 5: Marginal Product of Investment on Cognition Age $t + 1$



Notes: The figure shows marginal products of investment on cognition. The y-axis represents the impact on cognition, in standard deviation units, of increasing investment by one standard deviation. Panel (a) corresponds to the left panel of Figure 4(a) of AMN. Panels (b) and (c) are based on book expenditures in 100 INR and scaled test scores, respectively.

used to anchor investment. In the descriptive statistics in their Table 2, Attanasio et al. (2020) report expenditures in USD, but estimation is based on a standardized version of the expenditures in INR.⁹ Since the authors write “we use the same measurements at every age and we normalize on the amount spent on books” (p. 2524) and “our investment measure is measured in monetary units, reflecting cost” (p. 2526), it might not be clear to readers which units the results are based on. Ideally, scaling a measure in this way does not affect the main results, but I show that some of the main results are sensitive to using these standardized expenditures instead of book expenditures in 100 INR.¹⁰ Second the estimates in the paper are based on a standardized PPVT test score. I reestimate the model after multiplying all test scores by 3, which implies a variance of the scaled scores that is still well below the variance of the actual PPVT test scores. I then compare some of the estimates of Attanasio et al. (2020) with the ones obtained using the two scaled versions of the data.¹¹

Attanasio et al. (2020) analyze the marginal product of investment on cognition and

⁹While not discussed in their paper or in the code, their data set contains a standardized book expenditures variable, which turns out to be expenditures in INR, but standardized such that the mean and the variance of the pooled expenditures over all time periods are 0 and 1, respectively.

¹⁰This transformation amounts to multiplying the standardized test scores by 7.833 and adding 4.802.

¹¹In the first step, Attanasio et al. (2020) estimate the joint distribution of skills and then simulate from the estimated distribution to estimate production functions and to calculate counterfactuals. This first step of their code is sensitive to the starting values. To ensure that the results below are based on the same (local) minimum, I simply rescale the simulated variables instead of reestimating that part of the model.

write: “When considering the production of cognitive skills (...) the productivity of investments is much higher at younger ages: investments are more able to affect cognition earlier on.” This conclusion is based on the left panel of Figure 5 below, which corresponds to Figure 4(a) of Attanasio et al. (2020).¹² The middle panel of Figure 5 shows the results based on book expenditures in 100 INR. In this case, the marginal effects are higher at age 8 than at age 5 (and even negative in the latter case).¹³ When I rescale test scores instead of book expenditures, the estimated effects are much lower, as can be seen in panel (c).

By Theorem 4, there are different, invariant ways to interpret the effect of investment on skills. For example, here we can conclude: Fixing all variables at their median values,

- (i) if we increase someones investment from the 40% quantile to the 60% quantile at age 5, we increase her skills from the 46% quantile to the 51% quantile at age 8, or
- (ii) if we increase investment by 1% at age 5, we increase skills by 0.004% at age 8, or
- (iii) if we increase investment s.t. book expenditures increase by 1 std (or 783 INR) at age 5, we increase skills at age 8 s.t. the standardized PPVT score increases by 0.26 std (or 0.23 units).

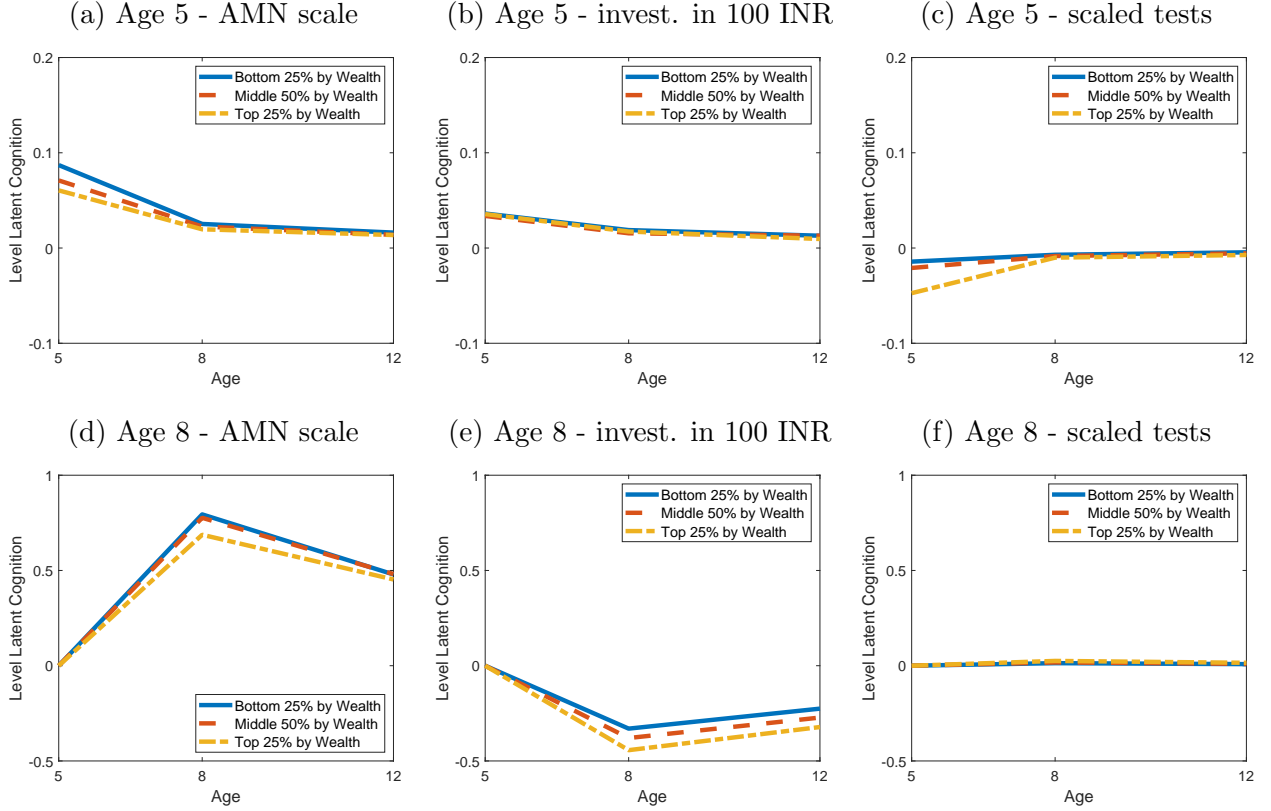
Although the effect described in interpretation (ii) may appear smaller than that in (i), the magnitudes are in fact consistent with each other. To understand their relationship, note that the logs of the latent variables follow mixtures of normal distributions. Here, the estimated variance of log-investment is substantially larger than that of log-skills. As a result, a 10% increase in investment at the median corresponds to a shift to the 50.18% quantile. In contrast, a 0.04% increase in skills at the median corresponds to a shift to the 50.07% quantile. The relative change in these quantiles ($0.0007/0.0018 = 0.39$), is comparable to the effect in interpretation (i), where the relative quantile change is $0.05/0.2 = 0.25$.

The authors also study the impact on cognition of an income transfer equal to 25% of mean income in the entire sample. Specifically, they consider the change in the median level of cognition in standard deviation units. The results for the different scales are shown in

¹²When calculating counterfactuals, they exclude the constant in the investment equation and incorrectly order some columns of the covariate matrix. I correct these issues and thus obtain slightly different estimates.

¹³I follow Attanasio et al. (2020) and do not restrict the signs of the coefficients (and many of their specifications result in some negative coefficients as well).

Figure 6: Estimated effects of income transfers - AMN estimator - skill level

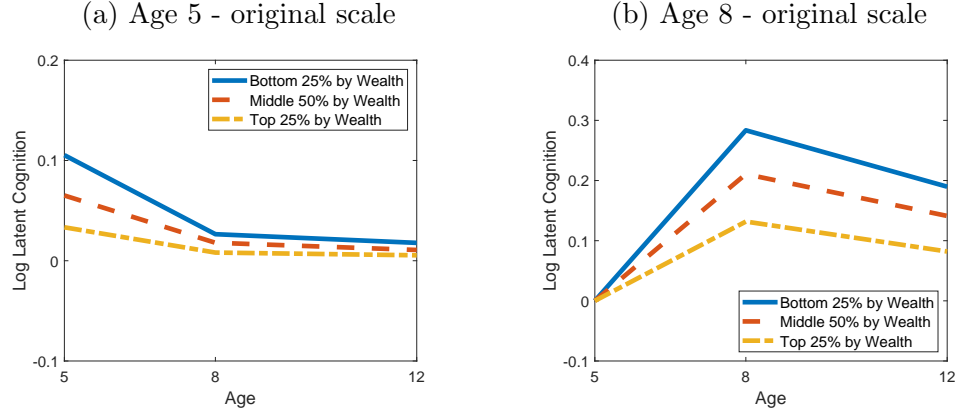


Notes: The figure shows the impact on cognition of an income transfer equal to 25% of mean income in the entire sample. The y-axis represents the impact on cognition, measured as the change in the median level of cognition in standard deviation units, of increasing investment by one standard deviation. In the top three graphs, the transfer is made before age 5. In the lower three graphs, it is made between ages 5 and 8. Panels (a) and (d) correspond to the left two panels of Figure 5 of AMN. Panels (b) and (e) are based on book expenditures in 100 INR. Panels (c) and (f) are based in scaled test scores.

Figure 6. The income transfer is made before age 5 in the upper panels and between ages 5 and 8 in the lower panels. The authors write: “In terms of timing, the largest impact is obtained if the transfer takes place when the children are between 5 and 8 for cognition”, which is based on the much larger response in the lower panel at age 12. This conclusion does not hold when using book expenditures in 100 INR, because the effect of investment is then estimated to be negative in panel (e). When scaling test scores instead, the effects are very close to 0 (see panels (c) and (f)).

There are two problems with the counterfactuals reported in Figure 6 using the specification of Attanasio et al. (2020). First, identified parameters are set to 1, which leads to misspecification. Second, this counterfactual does not correspond to those that are invariant

Figure 7: Estimated effects of income transfers - flexible estimator - log-skills



Notes: The figure shows the impact on cognition of an income transfer equal to 25% of mean income in the entire sample. The y-axis represents the impact on cognition, measured as the change in the median of the log of cognition in standard deviation units, of increasing investment by one standard deviation. In the left graph, the transfer is made before age 5. In the lower graph, it is made between ages 5 and 8. Estimates are based on the flexible estimator.

to the scale and location restrictions. To fix the second problem, one can focus on the change in the median log of skills rather than the level of skills.

The results for the log-skills using the estimator of Attanasio et al. (2020) are shown in Figure A.1 in Appendix E. Since the estimates are based on setting identified parameters to 1, the results are still sensitive to the scales. Figure A.2 in Appendix E shows the estimates corresponding to the counterfactual of Attanasio et al. (2020), but based on the estimator that also estimates the identified scales, which also depend on the the scales of the test scores (but not investment). Finally, Figure 7 displays the counterfactual in terms of log-skill based on the flexible estimator, which is invariant to the units of measurements (and therefore, the other scales are omitted). Based on Figure 7, we can still conclude that investment between ages 5 and 8 is more beneficial. However, compared to panels (d) of Figure 6, the new estimates imply much more heterogeneity by wealth. Figure A.3 in Appendix E shows the effects of income transfers on the entire counterfactual distribution of the test scores.

6 Conclusion

This paper demonstrates that in an important class of skill formation models, seemingly innocuous scale and location restrictions may not function as mere normalizations. Instead, they can constrain identified parameters and influence counterfactuals. The specific implications depend on the feature of interest, the production function, the measurement system, and the estimation method. In a new identification analysis, I pool all restrictions of the model and characterize the identified set without imposing additional scale and location restrictions. Notably, many key features remain invariant to these restrictions, are identified under weaker assumptions, yield robust policy implications, and are comparable across different studies.

Researchers often impose “normalizations” to achieve point identification in models that are otherwise only partially identified. To avoid unintended consequences and potentially misleading conclusions, it is critical to clarify which parameters and features are invariant to these restrictions, ensuring they are genuine normalizations as defined in Definition 1. While proving such invariance properties theoretically can be challenging, it is typically possible to demonstrate the robustness of key results to these restrictions.

One example closely related to the model considered here are skill formation models that explicitly solve household optimization problems to determine optimal investment functions (see e.g. Del Boca, Flinn, and Wiswall (2013) or Caucutt and Lochner (2020)). These models, which include a production function, are then estimated using the optimality conditions alongside a measurement system for latent variables. While these models have a different structure compared to the ones I study and a detailed analysis is left for future research, it appears that analogous normalization issues may arise, with consequences that depend on the model specification and the types of counterfactuals examined.

7 Data availability statement

The data and code underlying this research is available on Zenodo at <https://doi.org/10.5281/zenodo.16437164>.

References

- Adhvaryu, A., A. Nyshadham, and J. Tamayo (2023). Managerial Quality and Productivity Dynamics. *The Review of Economic Studies* 90(4), 1569–1607.
- Agostinelli, F. and M. Wiswall (2016a). Estimating the technology of children’s skill formation. NBER Working Paper No. 22442.
- Agostinelli, F. and M. Wiswall (2016b). Identification of dynamic latent factor models: The implications of re-normalization in a model of child development. NBER Working Paper No. 22441.
- Agostinelli, F. and M. Wiswall (2024). Estimating the technology of children’s skill formation. Accepted for publication at the *Journal of Political Economy*.
- Aguirregabiria, V. and J. Suzuki (2014). Identification and counterfactuals in dynamic models of market entry and exit. *Quantitative Marketing and Economics* 12, 267–304.
- Anderson, T. W. and H. Rubin (1956). Statistical inference in factor analysis. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 5: Contributions to Econometrics, Industrial Research, and Psychometry*, Berkeley, Calif., pp. 111–150. University of California Press.
- Attanasio, O., S. Cattan, E. Fitzsimons, C. Meghir, and M. Rubio-Codin (2020). Estimating the production function for human capital: Results from a randomized control trial in Colombia. *American Economic Review* 110(1), 48–85.
- Attanasio, O., S. Cattan, and C. Meghir (2022). Early childhood development, human capital, and poverty. *Annual Review of Economics* 14, 853–892.
- Attanasio, O., C. Meghir, and E. Nix (2020). Human Capital Development and Parental Investment in India. *The Review of Economic Studies* 87(6), 2511–2541.
- Attanasio, O., C. Meghir, E. Nix, and F. Salvati (2017). Human capital growth and poverty: Evidence from Ethiopia and Peru. *Review of Economic Dynamics* 25, 234–259.
- Aucejo, E. and J. James (2021). The path to college education: The role of math and verbal skills. *Journal of Political Economy* 129(10), 2905–2946.
- Bolt, U., E. French, J. Hentall-MacCuish, and C. O’Dea (2024). The intergenerational elasticity of earnings: Exploring the mechanisms. Working paper.
- Caucutt, E. M. and L. Lochner (2020). Early and late human capital investments, borrowing constraints, and the family. *Journal of Political Economy* 128(3), 1065–1147.
- Chiappori, P.-A., I. Komunjer, and D. Kristensen (2015). Nonparametric identification and estimation of transformation models. *Journal of Econometrics* 188(1), 22 – 39.
- Cunha, F. (2011). Recent developments in the identification and estimation of production functions of skills. *Fiscal Studies* 32(2), 297–316.
- Cunha, F. and J. Heckman (2007). The technology of skill formation. *American Economic Review* 97(2), 31–41.

- Cunha, F. and J. Heckman (2008). Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation. *Journal of Human Resources* 43(4), 738–782.
- Cunha, F. and J. Heckman (2009). The economics and psychology of inequality and human development. *Journal of the European Economic Association* 7(2), 320–364.
- Cunha, F., J. Heckman, and S. Schennach (2010). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica* 78(3), 883–931.
- Del Boca, D., C. Flinn, and M. Wiswall (2013, 10). Household choices and child development. *The Review of Economic Studies* 81(1), 137–185.
- Del Bono, E., J. Kinsler, and R. Pavan (2022). Identification of dynamic latent factor models of skill formation with translog production. *Journal of Applied Econometrics* 37(6), 1256–1265.
- Evdokimov, K. and H. White (2012). Some extensions of a lemma of kotlarski. *Econometric Theory* 28(4), 925–932.
- Fiorini, M. and M. Keane (2014). How the allocation of children’s time affects cognitive and noncognitive development. *Journal of Labor Economics* 32(4), 787–836.
- Freyberger, J. (2018). Nonparametric panel data models with interactive fixed effects. *Review of Economic Studies* 85(3), 1824–1851.
- Gallipoli, G. and S. Gomez (2023). The production of financial literacy. Working paper.
- Gao, W. and M. Li (2021). Robust semiparametric estimation in panel multinomial choice models. SSRN Working Paper No. 3282293.
- Hamilton, J., D. Waggoner, and T. Zha (2007). Normalization in econometrics. *Econometric Reviews* 26(2–4), 221–252.
- Heckman, J., R. Pinto, and P. Savelyev (2013, October). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review* 103(6), 2052–86.
- Helmert, C. and M. Patnam (2011). The formation and evolution of childhood skill acquisition: Evidence from India. *Journal of Development Economics* 95(2), 252–266.
- Hernández-Alava, M. and G. Popli (2017, Apr). Children’s development and parental input: Evidence from the UK millennium cohort study. *Demography* 54(2), 485–511.
- Kalouptzidi, M., P. T. Scott, and E. Souza-Rodrigues (2021). Identification of counterfactuals in dynamic discrete choice models. *Quantitative Economics* 12(2), 351–403.
- Komarova, T., F. Sanches, D. Silva Junior, and S. Srisuma (2018). Joint analysis of the discount factor and payoff parameters in dynamic discrete choice models. *Quantitative Economics* 9(3), 1153–1194.
- Lewbel, A. (2019). The identification zoo: Meanings of identification in econometrics. *Journal of Economic Literature* 57(4).
- Madansky, A. (1964). Instrumental variables in factor analysis. *Psychometrika* 29(2), 105–113.

- Matzkin, R. L. (1994). Restrictions of economic theory in nonparametric methods. Volume 4 of *Handbook of Econometrics*, Chapter 42, pp. 2523–2558. Elsevier.
- Matzkin, R. L. (2007). Nonparametric identification. Volume 6 of *Handbook of Econometrics*, Chapter 73, pp. 5307–5368. Elsevier.
- Murphy, K. M. and R. H. Topel (2016). Human capital investment, inequality, and economic growth. *Journal of Labor Economics* 34(2), 99–127.
- Rubio-Ramírez, J. F., D. F. Waggoner, and T. Zha (2010). Structural vector autoregressions: Theory of identification and algorithms for inference. *The Review of Economic Studies* 77(2), 665–696.
- Williams, B. (2020). Identification of the linear factor model. *Econometric Reviews* 39(1), 92–109.