A Simplified Klein–Spady Estimator for Binary Choice Models*

Per Hjertstrand† Andrew Proctor‡ Joakim Westerlund§

May 2025

Abstract

One of the most cited studies within the field of binary choice models is that of Klein and Spady (1993), in which the authors propose an estimator that is not only non-parametric with respect to the choice density but also asymptotically efficient. However, while theoretically appealing, the estimator has been found to be very difficult to implement with poor small-sample properties. This paper proposes a simplified version of the Klein–Spady estimator, which is shown to be easy to implement, numerically relatively more stable, and with excellent small-sample and asymptotic properties.

Key words: Binary choice; Maximum likelihood; Semi-parametric estimation.

1 Introduction

As Amemiya (1981) observed, binary choice models are "one of the most important developments in econometrics". This prominence stems largely from the widespread occurrence of

*Sofia Holmberg provided excellent research assistance. The computations for the Monte Carlo simulations in Section 3 were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at the PDC Center for High Performance Computing at the Royal Institute of Technology in Stockholm, and partially funded by the Swedish Research Council through grant agreement no. 2018-05973. Westerlund would like to thank the Knut and Alice Wallenberg Foundation for financial support through a Wallenberg Academy Fellowship.

†Research Institute of Industrial Economics (IFN), Sweden. E-mail: per.hjertstrand@ifn.se.

‡Department of Economics, Ludwig Maximilian University of Munich. E-mail: andrew.proctor@econ.lmu.de.

§Department of Economics, Lund University, Sweden, and Department of Economics, Deakin University, Australia. E-mail: joakim.westerlund@nek.lu.se.

binary choice settings across empirical research. Whether the focus is on individual behavior, institutional decisions, or market responses, many important questions reduce to choices between two alternatives: work or not, vote or abstain, adopt or reject a technology, and so on. Given how often such choices arise in applied contexts, credible estimation is essential for valid inference in social scientific research. This paper is consequently concerned with the rigorous estimation of binary choice models.

Let us therefore consider the binary variable $y_i \in \{0, 1\}$, observable for $i = 1, ..., n$ cross-sectional units. The data-generating process (DGP) that we will be considering for this variable is the same as in the bulk of the existing literature. We assume that the realization of y_i is the outcome of some latent continuous variable y_i^*, which can be interpreted as the utility difference between setting y_i to 0 or 1;

$$y_i := \mathbb{1}(y_i^* \geq 0), \tag{1}$$

where $a := b$ means that a is defined by b and $\mathbb{1}(A)$ is the indicator function of the event A being equal to 1 if A is true and 0 otherwise. We further assume that the latent y_i^* has a linear additive model representation:

$$y_i^* = x_i'\beta^0 + \varepsilon_i, \tag{2}$$

where $x_i \in \mathcal{X} \subseteq \mathbb{R}^r$ is a vector of observed regressors with $\beta^0 \in \mathbb{R}^r$ being a conformable vector of coefficients, and $\varepsilon_i \in \mathbb{R}$ is a stochastic error term. If we denote by $F(\cdot)$ the cumulative distribution function (CDF) of $(\varepsilon_i|x)$, then $\mathbb{P}(y_i = 1|x_i) = F(x_i'\beta^0)$, and since y_i is binary, we also have

$$\mathbb{E}(y_i|x_i) = F(x_i'\beta^0). \tag{3}$$

Because the conditional expectation is completely characterized as a function of $x_i'\beta^0$, the model is said to satisfy a "single-index restriction". It is useful to define $G(x_i'\beta^0) := \mathbb{E}(y_i|x_i)$ so as to keep this functional dependence explicit.

From this basic latent linear model framework, conventional estimators emerge based on the choice of the CDF $F(\cdot)$ of the error ε_i. Assuming that ε_i is either standard normally or logistically distributed leads to the probit and logit models, respectively. Meanwhile, the linear probability model (LPM) – which estimates a binary outcome using least squares linear

regression – corresponds to a latent linear model specification in which ε_i is uniformly distributed over a symmetric, bounded interval around zero (Delle Site and Parmar 2024).

These distributional assumptions on ε_i in the latent linear model are moreover foundational to the performance and relative merits of parametric binary choice estimators. This is explicitly the case for traditional binary choice estimators such as probit and logit, which are estimated via maximum likelihood (ML) and, as such, are consistent and efficient if the underlying assumption on the distribution of $F(\cdot)$ is correct, but otherwise are neither. Meanwhile, outside of the special case in which ε_i follows a symmetric uniform distribution, the LPM is well-known to be formally inconsistent and inefficient. Advocates of the LPM nevertheless prefer to use it because of its favorable properties as a linear approximator for nonlinear models (Moffitt 1999, and Angrist et al. 2006), accompanied by the popular but contentious heuristic belief that linear approximations might still function well in terms of the accuracy of the resulting inference (see, for example, Angrist and Pischke 2008, and Lewbel et al. 2012 for contrasting views). However, whether a linear approximation actually works well in binary choice settings once again hinges on the true distribution of $F(\cdot)$ (Aldrich and Nelson 1984).

The distribution of ε_i for the latent linear model is therefore critical to evaluating parametric binary choice models. The *assumed* distribution of ε_i gives rise to the various estimators, but equally the *actual* distribution of this error term determines whether or not these estimators perform well. As a result, much effort has gone into developing semi-parametric procedures that do not require correct specification of this distribution (see Ichimura and Todd 2007, for a comprehensive survey). One of the most well-cited studies within this strand of the literature is Klein and Spady (1993), henceforth abbreviated "KS". In this study, the authors propose a distribution-free approach to binary choice models that amends conventional parametric approaches by incorporating non-parametric estimation of $F(\cdot)$. The idea is to simply replace $F(\cdot)$ in the standard log-likelihood function with the Nadaraya–Watson estimator of the conditional expectation $G(\cdot)$, which by (3) is equal to $F(\cdot)$. The resulting estimator is not only $\sqrt{n}$-consistent and asymptotically normal, but also efficient in the sense that it attains the semi-parametric asymptotic efficiency bound that is analogous to the Cramér–Rao lower bound for parametric ML. As Mittelhammer and Judge (2011) point out, this last property makes the KS estimator a natural benchmark for binary choice problems.

However, while theoretically appealing, the KS estimator is difficult to implement, as the log-likelihood function can be badly behaved with saddle points and local optima. Many studies complain about this. For instance, Menezes-Filho et al. (2008, page 333) write: "We

choose polynomial estimation after finding the semi-parametric Klein and Spady (1993) estimator to exhibit problematic convergence behavior." Biewen et al. (2020, Footnote 5) similarly write: "In our empirical application, we initially experimented with the semi-parametric estimators developed by [...] Klein and Spady (1993). A disadvantage of these estimators is that they exhibit occasional convergence problems [...]." Rothe (2009), in work extending KS to allow for endogenous regressors, concludes: "One of the major issues of our estimator is its computational complexity when applied in settings with many regressors and/or observations. In this case, even evaluating the likelihood function at a specific point is very time consuming, and the function might have several local maxima. However, these problems are not specific to our [...] estimator but are encountered in general when computing semiparametric optimization estimators such as the ones by Ichimura (1993) or Klein and Spady (1993)" (pages 59–60). Discussions of this type extend even to econometrics textbooks – for example, Cameron and Trivedi (2005, page 485) state that "[t]he attraction of Klein and Spady's estimator is that it is fully efficient in the sense that it attains the semi-parametric efficiency bound. Computation is difficult, however." Due to these computational issues, a number of studies have found the estimator to perform poorly in small samples (see Chu et al. 2019, Frölich 2006, Mittelhammer and Judge 2011, Rothe 2009, and Westerlund and Hjertstrand 2014, to mention a few), while others indicate they had to exclude the KS estimator altogether (see, for example, Horowitz and Härdle 1996).

Observations like these have led to the development of alternative estimation approaches with improved numerical properties based on, for example, empirical likelihood functions, mixture distributions, local likelihood logit, indirect estimation, and Fourier flexible forms (see, for example, Coppejans 2001, Chen and Randall 1997, Frölich 2006, Mittelhammer and Judge 2011, and Westerlund and Hjertstrand 2014). However, these approaches are rarely used for several reasons. First, understanding and choosing between the various alternatives is technical and often opaque to many practitioners, with Ichimura and Todd (2007, page 5375) noting that "[a] barrier to implementing the new estimators is how to choose from a bewildering array of available estimators." Second, most such alternatives do not retain the same theoretical appeal in reaching the semi-parametric efficiency bound as with KS. Finally, they lack the same basic intuitive appeal of KS. In their popular textbook *Mostly Harmless Econometrics*, Angrist and Pischke (2008, page xii) write: "A principle that guides our discussion is that the estimators in common use almost always have a simple interpretation that is not heavily model dependent," citing the linear regression model as an exemplar for its simplic-

ity and robustness to common types of distributional misspecification. The promise of KS is much the same for binary choice settings; preserving the simple linear model structure for y_i^* that underlies standard parametric estimators, while eschewing the known $F(\cdot)$ assumption that is the central source of concern and disagreement among these estimators.

The current paper is motivated by recognition of both the strong appeal of KS as well as its downsides in practice. We propose a simplified version of KS, henceforth abbreviated "SKS", that retains the central approach and strengths of KS but amends the non-parametric estimation of $F(\cdot)$, such that the ubiquitous computational shortcomings arising with KS are practically eliminated. The way we accomplish this is by estimating the CDF $F(\cdot)$ directly via non-parametric kernel methods instead of estimating the conditional expectation $G(\cdot)$ via the Nadaraya–Watson estimator, which we argue is the source of many of the problems of the original KS estimator. In particular, unlike our proposed CDF estimator, the Nadaraya–Watson estimator is constructed as a ratio of estimated probability density functions (PDFs) that, in order to ensure asymptotic validity of the KS estimator, is not restricted to the unit interval. When this bound is exceeded, the single-index restriction (3) is violated, invalidating Nadaraya-Watson as an indirect estimator of $F(\cdot)$. Non-parametric PDF estimation is used in all fields of economics and statistics. As argued by Li and Racine (2007, page 23), the range of possible applications of non-parametric CDF estimation is equally as great, but it is not nearly as widely used. The current paper provides an example of a situation in which there is a clear preference towards CDF rather than PDF estimation. The SKS estimator that results from doing so is just as intuitive as the original. The main difference is that it is also fast, numerically stable, and has excellent small-sample properties.

The rest of this paper is organized as follows: In Section 2, we introduce the new estimator, and study its asymptotic properties. According to the results, SKS is not only $\sqrt{n}$-consistent and asymptotically normal, but also asymptotically efficient. These findings are supported by a large-scale Monte Carlo study, the results of which are reported in Section 3. The performance of SKS is compared to that of a number of other estimators, including KS. We confirm that while KS displays characteristic problems with convergence, and parametric estimators are heavily affected by distributional mispecification, SKS is both reliable and performant across a wide variety of DGPs. Indeed, SKS even displays near equal performance to probit when probit is the correct model. Section 4 is concerned with our empirical application to the behavioral determinants of savings behavior in youth. Using the same data as Sutter et al. (2013), we find that while KS once again shows numerical instability, SKS is numerically

very stable. The SKS results support the main finding of Sutter et al. – that impatience negatively affects the savings behavior of youth; however, their finding that math skills are also important for savings is not supported. We argue that this difference in the results is due to the preference of Sutter et al. to rely on probit even though the distribution of the regression errors is unknown and unlikely to be normal. Section 5 concludes. The paper is accompanied by an online appendix containing the complete set of Monte Carlo results, figures omitted from the paper, and some additional empirical results.

2 The SKS estimator and its asymptotic properties

2.1 The estimator

Suppose for a moment that $F(\cdot)$ is known. Since y_i is binary with $\mathbb{P}(y_i = 1|x_i) = F(x_i'\beta^0)$, we know that y_i is Bernoulli distributed with success probability $F(x_i'\beta^0)$. Thus, given $F(\cdot)$, the log-likelihood function is given simply by

$$\ell(\beta) := \sum_{i=1}^{n} [y_i \ln F(x_i'\beta) + (1 - y_i) \ln(1 - F(x_i'\beta))], \tag{4}$$

which when maximized leads to the ML estimator of β^0. Of course, this estimator is not feasible since in practice $F(\cdot)$ is unknown. Instead, KS suggest replacing $F(\cdot)$ with an estimate. Their choice of which estimator to use is based on the fact that the classical Nadaraya–Watson estimator is known to be consistent for the conditional expectation $G(x_i'\beta^0)$, which in view of (3) makes it consistent also for $F(x_i'\beta^0)$. In particular, KS propose using the following leave-one-out version of the Nadaraya–Watson estimator:

$$\hat{G}(x_i'\beta) := \frac{v(x_i'\beta)}{w(x_i'\beta)}, \tag{5}$$

where $v(x_i'\beta) := \sum_{j\neq i}^{n} k_h((x_j - x_i)'\beta)y_j$ and $w(x_i'\beta) := \sum_{j\neq i}^{n} k_h((x_j - x_i)'\beta)$ with $k_h(v) := k(v/h)$, $k(\cdot) : \mathbb{R} \to \mathbb{R}$ is a kernel function and $h > 0$ is a bandwidth parameter that may depend on n.[1] Replacing $F(\cdot)$ by $\hat{G}(\cdot)$ leads to the following feasible log-likelihood function:

$$\hat{\ell}_{\mathrm{KS}}(\beta) := \sum_{i=1}^{n} \tau_i [y_i \ln \hat{G}(x_i'\beta) + (1 - y_i) \ln(1 - \hat{G}(x_i'\beta))], \tag{6}$$

[1]The dependence of $\hat{G}(x_i'\beta)$, $v(x_i'\beta)$ and $w(x_i'\beta)$ on h is suppressed in order to avoid cluttering the notation.

where $\tau_i \in \mathbb{R}$ is a certain trimming term yet to be discussed. The KS estimator of β^0 is maximizer of $\hat{\ell}_{\text{KS}}(\beta)$.

In view of (5) and (6), it is clear that maximizing $\hat{\ell}_{\text{KS}}(\beta)$ can be difficult unless $w(x_i'\beta)$ – the denominator of $\hat{G}(x_i'\beta)$ – is bounded away from 0. This is an issue because in order to eliminate the asymptotic bias caused by the estimation of $G(x_i'\beta^0)$, KS requires the use of "higher-order" (or "bias-reducing") kernels with a certain number of zero moments.[2] But then these kernels are not even nonnegative. In fact, not only can $w(x_i'\beta)$ be zero, but $\hat{G}(x_i'\beta)$ as a whole is also not confined to lie between 0 and 1, which of course causes problems when taking logs. This is where the trimming term τ_i in (6) comes in. It is defined as $\tau_i := 1(x_i \in \mathcal{T})$, where $\mathcal{T} \subseteq \mathbb{R}^r$ is a certain compact set that is chosen such that the probability limit of $\hat{G}(\cdot)$ is bounded away from 0 and 1 on $\mathcal{T}$ (see, for example, Ichimura and Todd 2007, and Rothe 2009). The inclusion of τ_i in (6) takes care of the above mentioned problems but there are others. One additional problem is that the binary nature of y_i may lead to discontinuities in $\hat{G}(x_i'\beta)$, thus invalidating conventional gradient based optimization algorithms (see Westerlund and Hjertstrand 2014). Another problem is that while appealing from a theoretical point of view, higher-order kernels tend to suffer from very poor small-sample performance, to the point that many studies use standard kernels even though the resulting KS estimators are no longer asymptotically valid (see, for example, Rothe 2009).

The idea of the present paper is to avoid the aforementioned problems by estimating $F(x_i'\beta^0)$ directly instead of indirectly via (3) and estimation of $G(x_i'\beta^0)$. While there are many ways of doing this, in this paper we propose using non-parametric CDF estimation via kernel methods. Specifically, we suggest estimating $F(\cdot)$ using the following leave-one-out kernel CDF function (see, for example, Li and Racine 2007):

$$\hat{F}(x_i'\beta) := \frac{1}{n-1}\sum_{j\neq i}^{n} K_h((x_j - x_i)'\beta), \tag{7}$$

where $K_h(v) := K(v/h)$ and $K(v) := \int_{-\infty}^{v} k(w)dw$ is an integrated kernel. Unlike when $k(\cdot)$ is a high-order kernel, when it is standard second-order kernel $k(\cdot)$ is also a PDF. We will be working with second-order kernels, and therefore $k(\cdot)$ is a PDF, which in turn implies that $K(\cdot)$ is a CDF (see Li and Racine 2007). While $K(\cdot)$ can be chosen as any CDF, two natural choices arises by setting it equal to either the standard normal CDF, $K(v) = \Phi(v) :=$

[2]The order of a kernel is defined as the order of the first non-zero moment (see, for example, Li and Racine 2007, Chapter 1, for a formal definition and discussion).

$\int_{-\infty}^{v} e^{-w^2/2} dw/\sqrt{2\pi}$, or the logistic CDF, $K(v) = 1/(1+e^{-v})$, in which cases the SKS estimator can be interpreted as a "weighted" probit or logit estimator. Either way, the proposed SKS objective function is given by

$$\hat{\ell}_{\text{SKS}}(\beta) := \sum_{i=1}^{n} [y_i \ln \hat{F}(x_i'\beta) + (1-y_i) \ln(1-\hat{F}(x_i'\beta))], \tag{8}$$

and the resulting SKS estimator $\hat{\beta}$ of β^0 is defined as

$$\hat{\beta} := \arg\max_{\beta \in \mathcal{B}} \hat{\ell}_{\text{SKS}}(\beta), \tag{9}$$

where $\mathcal{B} \subseteq \mathbb{R}^r$ is the parameter space of β.

The main advantage of using $\hat{F}(\cdot)$ instead of $\hat{G}(\cdot)$ is that there is no need for trimming, since as already pointed out, $K(\cdot)$ is a CDF (see Berg and Politis 2009, for a discussion).

2.2 Asymptotic properties

We begin this section by stating the assumptions on which our asymptotic results are based. Because we are treating both β^0 and $F(\cdot)$ as unknown objects to be estimated from the data, we require a normalization to rule out degenerate cases. In particular, we know from Cosslett (1993) that without further restrictions, the constant term (if there is one) is not identified and that the slope coefficients are only identified up to scale. Let us therefore partition x_i as $x_i = (x_{1i}, x_{2i}')'$, where $x_{1i} \in \mathbb{R}$ is continuous and $x_{2i} \in \mathbb{R}^{r-1}$ is a vector containing all other regressors (with $r \geq 2$). The vector β^0 is partitioned conformably as $\beta^0 = (\beta_1^0, \beta_2^{0\prime})'$. The most common scale normalization scheme, which is also used in this study, is to set $\beta_1^0 = 1$ (see, for example, Ichimura 1993, KS and Rothe 2009). Location normalization is imposed by assuming that x_{2i} does not contain an intercept. The rest of the assumptions that we will be working under are stated below.

Assumption 1. (y_i, x_i) is independent and identically distributed across i.

Assumption 2. The parameter space $\mathcal{B}$ is compact and β^0 is an element of its interior.

Assumption 3. The support $\mathcal{X}$ of x_i is such that $0 < F(x_i'\beta^0) < 1$ for all i.

Assumption 4. The $r \times r$ matrix

$$\Sigma := \mathbb{E}\left[\frac{1}{F(x_1'\beta^0)[1-F(x_1'\beta^0)]}\frac{dF(x_1'\beta^0)}{d\beta}\left(\frac{dF(x_1'\beta^0)}{d\beta}\right)'\right]$$

is positive definite.

Assumptions 1–4 are standard in the literature on semi-parametric binary choice models (see, for example, Ichimura 1993, KS, Lee 1995, and Rothe 2009). We therefore do not comment on them.

Assumption 5. $F(x_i'\beta^0) = F(x_i'\beta)$ implies $\beta = \beta^0$.

Assumption 5 is also standard in the literature. It is the same as Assumption (C.9) in KS (see also Ichimura 1993, Lee 1995, and Rothe 2009). Sufficient conditions for Assumption 5 are provided in Section 2.2 of KS.

Assumption 6. The kernel function $k(\cdot)$ is continuously differentiable with the first derivative satisfying a Lipschitz condition. Also, $\int k(v)dv = 1$, $k(v) = k(-v)$, $\int v^2 k(v)dv > 0$ and $k(v) = 0$ for $|v| > 1$.

Assumption 6 allows for standard second-order kernels, which contrasts much of the previous literature where higher-order kernels are needed to reduce the error coming from the kernel estimation (see, for example, KS and Rothe 2009). As previously explained, the need for higher-order kernels is problematic because they are not strictly positive everywhere. Thus, estimators of the CDF based on higher-order kernel are not necessarily contained within the range $[0, 1]$ or restricted to be nondecreasing (see Berg and Politis 2009). These problems do not appear in the SKS estimator. In fact, as alluded to earlier in this section, under Assumption 6, $K(\cdot)$ satisfies all the conditions of a CDF.[3] This is an important advantage of the SKS estimator over any other estimator relying on higher-order kernels.

Assumption 7. The bandwidth h satisfies $h \sim n^{-\alpha}$ with $1/4 \leq \alpha < 1/3$.

KS use fourth-order kernels for constructing the Nadaraya–Watson estimator $\hat{G}(\cdot)$ of $G(\cdot)$. They require that $n^{-1/6} < h < n^{-1/8}$, which does not include the "optimal" rate that minimizes the asymptotic mean integrated squared error. For kernels of order ν, the optimal

[3]The only restriction in this regard is Assumption 3, which is again standard in the literature, and is not particularly restrictive. It holds if $\mathcal{X}$ is compact. Should Assumption 3 be deemed too restrictive, $\hat{F}(\cdot)$ can be trimmed without cost (see Berg and Politis 2009).

bandwidth is proportional to $n^{-1/(2\nu+1)}$ (see, for example, Li and Racine 2007, Chapter 1). The optimal rate for $\nu = 4$ is therefore given by $n^{-1/9}$, which is outside the permissible range of KS. The same is true here. However, there is an important difference. Unlike KS, we estimate $F(\cdot)$, as opposed to $G(\cdot)$, and the optimal bandwidth for CDF estimators is proportional to $n^{-1/3}$ (see Hansen 2004, or Li and Racine 2007, Chapter 1). Although $n^{-1/3}$ is outside the permissible range given in Assumption 7, it nevertheless comes arbitrary close to satisfying Assumption 7. Indeed, h can be made arbitrarily close to optimal by setting it proportional to $n^{-1/(3+\epsilon)}$, where ϵ is an arbitrary small positive number.[4]

We now have all the assumptions we need to state our first asymptotic result.

Theorem 1 (Consistency)**.** *Suppose that Assumptions 1–3 and 5–7 are met. Then, as $n \to \infty$,*

$$\hat{\beta} \to_p \beta^0.$$

Having established the consistency of the new estimator, we now turn to its asymptotic distribution. The route to asymptotic normality standardly involves use of the mean value theorem, and the convergence in probability of the Hessian in a neighborhood of β^0 (see, for example, KS, Lee 1995, and Rothe 2009).

We begin by applying the mean value theorem to $d\hat{\ell}_{\text{SKS}}(\hat{\beta})/d\beta$ around $\hat{\beta} = \beta^0$. This gives

$$0_{r\times 1} = \frac{d\hat{\ell}_{\text{SKS}}(\hat{\beta})}{d\beta} = \frac{d\hat{\ell}_{\text{SKS}}(\beta^0)}{d\beta} + \frac{d^2\hat{\ell}_{\text{SKS}}(\beta^*)}{d\beta(d\beta)'}(\hat{\beta} - \beta^0), \tag{10}$$

where β^* lies element-wise between the line segment joining $\hat{\beta}$ and β^0. By solving this equation for $\sqrt{n}(\hat{\beta} - \beta^0)$, we obtain

$$\sqrt{n}(\hat{\beta} - \beta^0) = \left(-\frac{1}{n}\frac{d^2\hat{\ell}_{\text{SKS}}(\beta^*)}{d\beta(d\beta)'} \right)^{-1} \frac{1}{\sqrt{n}}\frac{d\hat{\ell}_{\text{SKS}}(\beta^0)}{d\beta}. \tag{11}$$

Lemmas 1 and 2 below provides the limiting behaviour of the components of (11).

[4]Since $\epsilon > 0$ can be made arbitrarily close to zero, in empirical implementations of the SKS estimator it makes no difference to actually set $\epsilon = 0$ and choose $h \sim n^{-1/3}$. In the Monte Carlo experiments and empirical applications in Sections 3 and 4, we therefore set $h = n^{-1/3}$.

Lemma 1 (Hessian)**.** *Suppose that Assumptions 1–7 are met. Then, as* $n \to \infty$*,*

$$-\frac{1}{n}\frac{d^2\hat{\ell}_{\text{SKS}}(\beta^*)}{d\beta(d\beta)'} \to_p \Sigma.$$

Lemma 2 (Score)**.** *Suppose that the conditions of Lemma 1 are met. Then, as* $n \to \infty$*,*

$$\frac{1}{\sqrt{n}}\frac{d\hat{\ell}_{\text{SKS}}(\beta^0)}{d\beta} \to_d N(0_{r\times 1}, \Sigma).$$

The asymptotic distribution of $\sqrt{n}(\hat{\beta} - \beta^0)$ is an immediate consequence of (11), Lemmas 1 and 2, and Assumption 4, and is given in Theorem 2 below.

Theorem 2 (Asymptotic distribution)**.** *Suppose that the conditions of Lemma 1 are met. Then, as* $n \to \infty$*,*

$$\sqrt{n}(\hat{\beta} - \beta^0) \to_d N(0_{r\times 1}, \Sigma).$$

Similarly to KS and Ichimura (1993), estimation of the covariance matrix Σ can be performed in the usual way. The SKS objective function can be treated as though it were the true log-likelihood function, and standard errors obtained from a conventional likelihood routine will be asymptotically correct. It is important to note that inference based on such standard errors is robust to multiplicative heteroscedasticity of a general but known form and heteroscedasticity of an unknown form if it depends only on the single-index restriction (see KS for a discussion).

KS (Theorem 5) showed that their estimator attains the semi-parametric efficiency bound. Because the asymptotic covariance matrix given in Theorem 2 is the same as the one given in Theorem 4 of KS, the proposed SKS estimator attains the same bound. It is therefore asymptotically efficient. Theorem 3 formalizes this.

Theorem 3 (Asymptotic efficiency)**.** *Suppose that the conditions of Lemma 1 are met and that* x_i *and* ε_i *are independent. Then,* $\hat{\beta}$ *is asymptotically efficient.*

3 Monte Carlo simulations

A large-scale Monte Carlo study was conducted to evaluate the small-sample properties of the SKS estimator relative to KS, probit and the LPM. The full set of results is too numerous to report here in full. Consequently, in this section, we report aggregated results and present all disaggregated results in Sections A and B of the online appendix.

3.1 Setup

The DGP is given by a restricted version of (1) and (2) that sets $x_i = (1, x_i^c, x_i^b)'$, where $x_i^c \in \mathbb{R}^{r^c}$ and $x_i^b \in \mathbb{R}^{r^b}$ are vectors continuous and binary regressors, respectively. The associated vector of coefficients is given by $\beta^0 = (\alpha, \beta_1^0, \beta_2^{0\prime})'$, where $\beta_1^0 = -1$ is the slope of the first continuous regressor, while $\beta_2^0 = (1, -1, 1, ...)'$ contains the slopes of the remaining $r - 2$ regressors. This decomposition of β^0 will be useful later. Similarly to Westerlund and Hjertstrand (2014), the intercept, α, is calibrated so that $\overline{y} = \sum_{i=1}^n y_i/n$ is "close" to the targeted success probability, which is one of 0.25, 0.5 and 0.85.[5] Following Frölich (2006), the following four regressor designs are considered:

R1. The first design contains one continuous and one binary regressor ($r^c = r^b = 1$). The continuous regressor is drawn from a chi-square distribution with one degree of freedom, henceforth denoted $\chi^2(1)$, while the binary regressor is drawn from a Bernoulli distribution with survivor intensity set to 0.5, henceforth denoted $\text{Ber}(0.5)$.

R2. In this design, $r^c = 5$ and $r^b = 1$. The first continuous regressor, $x_{1,i}^c$ say, is drawn from $\chi^2(1)$, while the remaining four, $x_{2,i}^c, ..., x_{5,i}^c$, are generated as $x_{j,i}^c \sim x_{j-1,i}^c + \chi^2(1)$ for $j = 2, ..., 5$. As in R1, the binary regressor is drawn from $\text{Ber}(0.5)$. The implied correlation among the continuous regressors is between 0.5 and 0.9.

R3. Now, $r^c = 1$ and $r^b = 5$. The first binary regressor, $x_{1,i}^b$ say, is drawn from $\text{Ber}(0.5)$, while the remaining four are generated as $x_{j,i}^b \sim \text{Ber}(0.4 + 0.4\overline{x}_{j-1,i}^b)$, where $\overline{x}_{j,i}^b = \sum_{l=1}^j x_{l,i}^b/j$. Hence, if all preceding regressors are equal to 1, the probability that the next regressor takes on the same value one is 0.8. As in R1, the continuous regressor is drawn as $\chi^2(1)$. This means that the correlation between the binary regressors is between 0.1 and 0.4.

[5] $\overline{y}$ is calibrated to be at most 0.025 away from the targeted success probability in at least 80% of all replications. Figure C1 in Section C of the online appendix gives a histogram of the absolute difference between $\overline{y}$ and the desired probability over all replications in all Monte Carlo exercises.

R4. In the last design, $r^c = r^b = 5$ with the continuous and binary regressors generated as in R2 and R3, respectively.

For the error term, ε_i, we consider another five designs, which are largely based on Frölich (2006):

E1. In the first design, $\varepsilon_i \sim N(0,1)$, which means that the probit model is correctly specified.

E2. In the second design, ε_i is drawn from a t-distribution with one degree of freedom, henceforth denoted $t(1)$. This is a symmetric and heavy-tailed distribution where the mean and variance are undefined, but where the median is 0.

E3. In this design, $\varepsilon_i \sim \chi^2(3) - 3$, which means that the distribution of ε_i is skewed with mean 0 and variance 6.

E4. Here, $\varepsilon_i = b_i(l_i + 2) + (1 - b_i)(l_i - 2)$, where $b_i \sim \text{Ber}(0.5)$ and l_i is a Laplace distributed variable with location and scale parameters set to 0 and $1/\sqrt{2}$, respectively. This implies that the distribution of ε_i is bimodal with mean 0 and variance 5.

E5. In the fifth and final design, $\varepsilon_i \sim t(2) \cdot 0.14\sqrt{\sum_{k=1}^{r^c}\sum_{j=1}^{r^b} x_{k,i}^c x_{j,i}^b}$. The distribution of ε_i is therefore heteroskedastic with mean 0 and infinite variance.

For each of the four regressor designs, five error designs and three success probabilities, we consider three sample sizes; $n \in \{250, 500, 1500\}$ (similarly to Rothe 2009). This resulted in a total of 180 parameterizations of the DGP. The number of replications in each experiment is set to 1000.

3.2 Implementation

The SKS estimator is implemented as described in Section 2, with $K(\cdot)$ set equal to the logistic CDF.[6] As discussed in Section 2, in the estimation β^0 is normalized by removing the intercept and normalizing the coefficient of the first continuous regressor, β_1^0, to 1. The bandwidth h is set equal to $n^{-1/3}$, which, as explained in Section 2, is proportional to the optimal choice.

In interest of comparison, we also simulate the LPM, probit and three versions of the KS estimator. The LPM and probit are fitted with an intercept. The KS estimator is implemented

[6]We also experimented with $K(\cdot)$ set to the normal CDF. However, we found the logistic CDF to be more numerically stable than the normal, and therefore opted for this choice.

as described in Section 2 with $k(\cdot)$ set equal to the logistic PDF. We impose the same normalization as for the SKS estimator. The three KS versions differ only in the choices of bandwidth. The first version, labelled "KS1", is based on setting $h = n^{-1/6.02}$, just as in the Monte Carlo study of KS. The second version, labelled "KS2", treats h as an additional parameter that is estimated along with the slope coefficients (as in, for example, Frölich 2006, and Rothe 2009).[7] In the third version, h is determined based on generalized cross-validation (similarly to, for example, Newey et al. 1990, Gerfin 1996, and Frölich 2006), which is implemented as outlined in Section 5 of Delecroix et al. (2006). However, since the performance of this version turned out to be considerably worse than for KS1 and KS2, we do not present the results.[8]

In order to assess the numerical stability of the estimators, we consider multiple starting values (similarly to Westerlund and Hjertstrand 2014). Let us therefore denote by β_1^s and β_2^s the starting values for β_1^0 and β_2^0, respectively. The following six sets of starting values are considered:

S1. The starting values are set to the true coefficients; that is, $\beta_1^s = \beta_1^0$ and $\beta_2^s = \beta_2^0$.

S2. β_1^s and β_2^s are set equal to their corresponding LPM estimates.

S3. β_1^s and β_2^s are set equal to their corresponding probit estimates.

S4. $\beta_1^s = |\beta_1^0|$ and $\beta_2^s = |\beta_2^0|$.

S5. $\beta_1^s = |\beta_1^0|$ and $\beta_2^s = 2.5|\beta_2^0|$.

S6. $\beta_1^s = |\beta_1^0|$ and $\beta_2^s = 5|\beta_2^0|$.

For KS2, the starting value for h is always set to $n^{-1/6.02}$. For the probit estimator, we naturally drop S3. The starting value for the constant is set to its true value, α. The LPM obviously does not require initialization and is therefore only presented once.

[7] Härdle et al. (1993) propose estimating h together with the coefficients in semi-parametric single-index models using a weighted least squares approach. Although it does not seem to be theoretically verified, it is widely conjectured that the same approach can be used together with the KS estimator.

[8] These results are available upon request from the authors. In our simulations, we did not record a single instance where the KS estimator based on generalized cross-validation performed better than KS1 and KS2 in terms of bias and root mean squared error (RMSE). In fact, the difference in performance was in a large majority of the Monte Carlo experiments 10–100 times worse. The poor performance of generalized cross-validation is consistent with the working paper version of Delecroix et al. (2006). On this issue, they say that [brackets added] "repeating Steps I.1 and I.2 until convergence one expects some values [of the parameters] very close to those obtained by joint estimation [of the parameters and bandwidth (as in KS2)]. However our experience proves that iterating Steps I.1 and I.2 is not only more computational demanding but also leads to more instable results."

All computational work was carried out in Matlab. For SKS, KS1 and KS2, the log-likelihood functions were optimized using the routine `fminunc` from the optimization toolbox. Only default settings were used with user-specified expressions for the gradients and Hessian, where Fisher scoring algorithms were used to maximize the log-likelihood functions. The probit estimator was implemented using the `glmfit` command from the Statistics and Machine Learning toolbox.[9]

3.3 Results

The accuracy of the estimated slope coefficients is assessed based on their bias, root mean squared error (RMSE) and median absolute deviation (MAD). As already pointed out, for each estimator we consider 180 parameterizations of the DGP, and for KS1, KS2, KS3, and SKS we also consider six sets of starting values (for probit we consider five starting values). This means that the total number of constellations of DGPs, estimators and stating values is no less than 4500.[10] Therefore, in order to report our findings succinctly, in this section we follow Frölich (2006) and aggregate the results across regressor designs, error designs and success probabilities.[11] The complete disaggregated results can be found in Sections A and B of the online appendix.

INSERT TABLES 1 AND 2 ABOUT HERE

Tables 1 and 2 present the results for each specification of starting values and n. While Table 1 contains the results for E1, Table 2 reports the results aggregated over all other error designs. The reason for presenting the results in this way is that while in E1 probit is efficient as ε_i is normally distributed, in E2–E5 it is inconsistent. The information content of Tables 1 and 2 can be summarized in the following way:

- As expected, probit performs well in E1 but not in E2–E5. Moreover, while generally decreasing in n in E1, the probit RMSE is increasing in n in E2–E5, which is a reflection

[9] The simulated data and codes to replicate all Monte Carlo results are available from the authors upon request.

[10] While not a problem in single applications to real data, with 1000 replications of 4500 DGP constellations – some of which are rather extreme – occasionally estimation will break down. As a way to address this problem, we trim out (estimated) success probabilities close to 0 or 1, and only keep pairs (y_i, x_i) for which this probability is in the interval $[\Delta, 1-\Delta]$. Consistent with the findings of studies such as KS, Lee (1995), and Rothe (2009), the choice of trimming threshold Δ did not matter much. We therefore set it arbitrarily to $\Delta = 10^{-12}$. This is done for all estimators but probit and the LPM.

[11] In particular, we stack every Monte Carlo replication in a single vector implying that each replication is given the same weight in the aggregated results.

of the fact that the estimator is consistent only in E1.

- The LPM performs reasonably well, and it does so regardless of the error design considered. Since no starting values are required, it is also numerically stable.
- KS1 is generally numerically more stable than the computationally more costly KS2.
- Initializing KS1, KS2 and SKS at the probit estimates (S3) generally results in poor RMSE performance in E2–E5, which is unsurprising given the poor performance of probit. LPM initialization (S2) seems to work much better in this regard.
- SKS is the best-performing estimator in almost every Monte Carlo experiment, with RMSE and MAD values that are often several times lower than those of KS1 and KS2. SKS is also numerically stable in the sense that it mostly converges to the same solution for different starting values. It is only when the initialization is carried out using probit that the RMSE performance of SKS is poor.
- SKS performs well even when compared to probit in E1, which is noteworthy since in this design, probit is again based on the true error distribution.

All in all, we find that the proposed SKS approach generally performs well in small samples and across the DGPs considered. SKS not only outperforms KS1 and KS2 in almost every Monte Carlo experiment, but it also performs comparably to probit when probit is the correctly specified model. This combination of performance, robustness, and numerical stability makes it especially appealing for empirical work. We therefore believe that SKS merits serious consideration among the bevy of estimators for binary choice models.

4 Empirical application

4.1 Background and data

In this section, we apply our proposed SKS estimator to analyze the economic behavior of youth. During the last two decades, the economic decision making of children and adolescents has attracted considerable attention, so much that there is by now a separate literature devoted to it (see Sutter et al. 2019, for a recent survey). Experimental elicitation of behavioral economic preferences – chiefly risk and time parameters – combined with observation of economic behavior in the field suggests a strong correlation between these parameters and

fundamental behavior both during youth and later adulthood, suggesting the possibility that these parameters have a fundamental role in shaping life outcomes.

Sutter et al. (2013) were among the first to document this relationship between experimentally derived preferences and behavior outside the lab for youth. They conducted experiments with 661 children and adolescents, aged 10 to 18 years, through which they obtained measures of impatience, risk aversion, and ambiguity aversion. They then regressed these experimental measures on five measures of behavior in the field, including saving, health, and school performance. Except for their health measure, the dependent variables are all binary. The reported results support the idea that children's and adolescents' experimental choices are related to their field behavior.

We examine this study as an instance of common practice in applied research, where the binary choice specification (probit in this study) is made without providing any justification.[12] The worry, of course, is that the parametric specification assumes the errors follow a given distribution, which may not be true. In the case of Sutter et al., the probit model specification implies an assumption that the errors are normally distributed. If normality is violated, then, as pointed out earlier, probit is inconsistent, casting doubt on the results reported by Sutter et al., an issue that, to the best of our knowledge, has not been considered before. In contrast to a parametric approach, such as probit, KS and SKS are asymptotically valid even when the distribution of the stochastic errors is unknown. They are therefore more robust in this regard. The purpose of this section is to investigate the extent to which the conclusions of Sutter et al. hold up when the estimation is carried out semi-parametrically using the KS and SKS estimators.

The data we use are taken directly from Sutter et al..[13] In their study, they perform a number of analyses, based on having multiple outcomes and several alternative measures of impatience. In order to provide a detailed comparison, we focus on one of these specifications,

[12] We emphasize, however, that this is by no means a criticism of the authors. The vast majority of binary choice estimation in empirical work employs parametric specifications, nearly always without any distributional justification. We find this unsurprising as we believe it is unlikely that most contexts provide a compelling ex ante justification for an imposed distributional assumption. Rather than attempt to justify the distributional assumption, the most common approach to addressing the risk of distributional mispecification is to estimate multiple parametric models (for example, both the probit and linear probability models). We emphasize that this is also problematic, however, because not only is it unclear that any of the parametric specifications are reasonable in a given setting, but inference only when results are similar or uniformly statistically significant across multiple models, some of which are wrong by construction, will generate uncontrolled distortions in both Type I and Type II error probabilities.

[13] The data can be downloaded from the web site of the American Economic Review at `https://www.aeaweb.org/articles?id=10.1257/aer.103.1.510`.

which takes an indicator for savings behavior as the outcome, and estimate separate models for each measure of impatience ("Models A-H" in Sutter et al.). Since the results for all models are very similar, and the same conclusions can be drawn independently of the model, we focus our discussion on the model that measures impatience based on low-stakes choices between immediate returns and returns with a 3-week delay ("Model A" in Sutter et al.). The results from Models B-H using all other different measures of impatience are presented in Section D of the online appendix. Aside from impatience, the specification includes as regressors measures of risk aversion and ambiguity aversion, age, an indicator for whether the participant is female, grades in German and math, number of siblings, and the participant's weekly amount of pocket money.

4.2 Implementation

We employ the same LPM, probit, KS1, KS2 and SKS estimators as in the Monte Carlo study of Section 3. The KS estimator is available as a built-in command in several statistical softwares and the details of the implementation varies (as opposed to the LPM and probit). In order to assess the robustness of our conclusions in this regard, in addition to the above mentioned estimators, we also report the results obtained by applying KS using Stata's `sml` command (see DeLuca 2008), henceforth referred to as "KS3". We therefore consider a total of six estimators in this section. For each estimator, we consider multiple starting values. In particular, in addition to S2 and S3 (the LPM and probit) we also consider feasible versions of S5 and S6, henceforth denoted "S5′" and "S6′", respectively, in which $\beta_1^s = \hat{\beta}_{1,\text{LPM}}$ and $\beta_2^s = \kappa|\hat{\beta}_{2,\text{LPM}}|$, where $\hat{\beta}_{1,\text{LPM}}$ and $\hat{\beta}_{2,\text{LPM}}$ are the LPM estimates of β_1^0 and β_2^0, respectively, $\kappa = 2.5$ in S5′ and $\kappa = 5$ in S6′. KS1–KS3 and SKS are initiated using all four specifications. Probit is for obvious reasons only initiated using S2, S5′ and S6′. As in Section 3, $k(\cdot)$ is chosen to be the logistic PDF.

As in Section 2, we require the normalization of one coefficient to 1 for identification. We choose to normalize upon the impatience parameter, both so that we may compare other estimates relative to this main parameter of interest in Sutter et al. (2013), but also because this is the only variable for which the estimate is clearly different from 0 (see Rothe 2009). Because the estimated effect is negative, the sign of the normalized coefficients change. The normalization of the LPM and probit models were carried out post-estimation, which means that their standard errors cannot be obtained in the usual manner, but must instead be obtained via the Delta method. The Delta method was implemented using both robust and non-robust

standard errors from the probit and LPM estimators. Since the robust and non-robust standard errors only marginally differ, with no change in parameter significance, we only present results based on implementing the Delta method using non-robust standard errors. The standard errors of the KS and SKS estimators are calculated from the inverse of the Fisher information matrix. Recall that inference in the KS and SKS estimators is robust to multiplicative heteroscedasticity of a general but known form and heteroscedasticity of an unknown form if it depends only on the single-index restriction.

In addition to the estimated coefficients, we calculate marginal effects. For KS, these effects are given by:

$$\frac{d\hat{G}(x_i'\beta)}{dx_i} = \hat{g}(x_i'\beta)\beta := \left[-\frac{1}{h} \frac{\frac{dv(x_i'\beta)}{dx_i'\beta} w(x_i'\beta) - v(x_i'\beta)\frac{dw(x_i'\beta)}{dx_i'\beta}}{w(x_i'\beta)^2} \right] \beta. \tag{12}$$

An important point about (12) is that $\hat{g}(\cdot)$ may be negative, implying that marginal effects may have different signs than β. If this is the case, $\hat{g}(\cdot)$ is not a PDF, which invalidates $d\hat{G}(x_i'\beta)/dx_i$ as a measure of marginal effects.[14] For the data at hand, $\hat{g}(\cdot)$ is negative in 77% of all observations, suggesting this is a nontrivial problem.

The marginal effects for the SKS estimator are given by

$$\frac{d\hat{F}(x_i'\beta)}{dx_i} = \hat{f}(x_i'\beta)\beta := \left[\frac{1}{(n-1)\,h} \sum_{j \neq i}^{n} k_h((x_j - x_i)'\beta) \right] \beta, \tag{13}$$

where $\hat{f}(\cdot)$ is the standard (leave-one-out) kernel PDF function. In contrast to $\hat{g}(\cdot)$, $\hat{f}(\cdot)$ is nonnegative by construction regardless of the choice of $k(\cdot)$, provided of course that it is a valid kernel function. Thus, unlike $d\hat{G}(x_i'\beta)/dx_i$, $d\hat{F}(x_i'\beta)/dx_i$ always has the same (correct) sign as β.

Both (12) and (13) are based on setting the bandwidth equal to $h = (n\pi^4/63)^{-1/5}$, which minimizes the integrated mean squared error of $\hat{f}(\cdot)$ when $k(\cdot)$ is the logistic PDF (Abo-El-Hadid 2018). Hence, unlike for the coefficients, for the marginal effects we only compute one version of KS (and SKS).

[14] This issue arises because the Nadaraya–Watson estimator, $\hat{G}(x_i'\beta)$, is not bounded away from 0 as explained in Section 2.

4.3 Results

Tables 3 and 4 report the estimated coefficients. While Table 3 contains the results for KS1–KS3, Table 4 contains the results for the LPM, probit and SKS. KS2 (where the bandwidth is estimated jointly with the coefficients) failed to converge when initialized based on S5′. The same problem occurred for probit when initialized based on S5′ and S6′. The results for these specifications are therefore omitted.

INSERT TABLES 3 AND 4 ABOUT HERE

The first thing to note about Table 3 is that all three versions of KS are highly sensitive to the choice of starting values. This is visible from the KS1 and KS2 implementations that use Matlab, but it is even more apparent from the KS3 implementation using the `sml` command in Stata. Indeed, looking across the four initializations, we see that in absolute value terms, the estimated coefficients can be several times larger for one initialization than for another, and that many coefficient estimates even change signs. The significance of the estimated coefficients varies, too. Many estimates are highly significant for some initializations and insignificant for others. This sensitivity is problematic not only from a reliability point of view but also because it lends itself to misuse, as researchers may obtain almost any result they want by a creative choice of initialization.

The picture is quite different if we instead look at the SKS estimates reported in Table 4. In fact, at the third decimal level of accuracy considered in the table, the estimation results do not depend on the initialization at all. This suggests that SKS is able to locate the global optimum and that the SKS objective function is considerably more well-behaved than the KS objective function. The standard errors also do not change and as a result the significance of the estimated coefficients is unaffected by the initialization.

INSERT TABLE 5 AND FIGURE 1 ABOUT HERE

In Table 5, we further contrast the properties of the SKS and KS estimators by examining the distribution of their estimated marginal effects for every regressor over the full sample.[15] Consistent with the results for the estimated coefficients, we see that for SKS the marginal effects do not differ much across observations, and that they have the same (correct) sign as the corresponding coefficients. By contrast, for KS the marginal effects vary quite substantially

[15] We treat all regressors as continuous because of the difficulty to calculate marginal effects for binary variables for semi-parametric kernel estimators as the one considered in this paper.

and many change sign. The reason for the change of signs is, as already pointed out, that $\hat{g}(x_i'\beta)$ may be negative. In order to illustrate this point, in Figure 1 we plot $\hat{g}(x_i'\beta)$ and $\hat{f}(x_i'\beta)$ over the values of $x_i'\beta$. Two observations stand out. First, unlike $\hat{f}(\cdot)$, $\hat{g}(\cdot)$ has several kinks and flat segments, which in view of (12) suggests that gradient-based optimization of the KS objective function can be difficult. Second, $\hat{g}(\cdot)$ is negative for a majority of values.

Because of the above results, we hereafter disregard KS. Comparing instead the SKS results to those obtained by the LPM and probit (Table 4), we first see that although the results tend to be directionally consistent across specifications, the precise point estimates vary by estimator. From a significance standpoint, all three approaches agree that age is a significant predictor of savings behavior, while risk and ambiguity aversion, gender, grades in German, and number of siblings appear to have only small, insignificant relative effects on savings. The main difference between SKS, on the one hand, and the LPM and probit, on the other hand, concerns the effect of math grades, which is large and statistically significant according to the LPM and probit, but small and insignificant according to SKS.

It is worth considering what this overall set of results may indicate about the comparison of parametric and semiparametric estimation in the Sutter et al. setting. First, the predominant similarity of the parametric results to the SKS results suggests that, generally, parametric estimators yield reasonable approximations of the preferred semiparametric model for this setting. Yet the agreement is not unequivocal, and where the two approaches differ in their findings (for math grades), the SKS should likely be preferred: although the parametric approximations appear generally quite close to the semiparametric model, the model errors don't appear to be either standard normal or uniform, in which case the parametric estimators considered remain inconsistent for the structural parameters.

Still, there is the question of why inference for math grades in particular might differ, given that we find the parametric estimators yield generally reasonable approximations of the semiparametric model for this setting. Considering especially the formal inconsistency of the parametric estimators when the error distribution is misspecified, this discrepancy might arise for several reasons. But one possibility that we wish to highlight here speaks to the broader literature around math skills, impatience, and savings.

In this literature, on the one hand a number of studies attest to a strong correlation between numeracy or math skills and positive financial behaviors, including savings (e.g. Banks and Oldfield 2007; Benjamin et al. 2013; Brounen et al. 2016; Estrada-Mejia et al. 2016; Lusardi 2012; Marley-Payne et al. 2022), with some quasi-experimental studies further supporting a

positive causal effect of math skills on financial behavior in some settings (Bernheim et al. 2001; Brown et al. 2016; Cole et al. 2016). Yet a number of other studies also document correlation between time preferences and educational performance (e.g. Cadena and Keys 2015; Castillo et al. 2019; Delaney et al. 2013; Golsteyn et al. 2014; Hanushek et al. 2023, 2021; Horn and Kiss 2020, and specifically math performance (Castillo et al. 2011; Lührmann et al. 2018). We similarly find that math grades are a significant predictor of impatience in the Sutter et al. (2013) sample.

In the context of this relationship between math grades and impatience, one possible explanation for the significance of math grades in probit and LPM is that the inconsistency of parametric models may lead to misleading significance of irrelevant regressors ("math grades") that are significant predictors of relevant included regressors ("impatience"). The reason for this is that the misspecification error that results from an erroneous parametric assumption enters as part of the regression error and depends on the relevant regressors of the model. Consequently, irrelevant regressors may become significant because of their correlation with the error term. Thus, although math skills may indeed have an independent effect on savings in some settings, it is possible that the reason why parametric estimation yields significant estimates of math grades in this setting – unlike SKS and in contrast to their typical agreement – is instead due to the well-known relationship between math grades and impatience in the context of a misspecified parametric model.

Summarizing, the predominant similarity of results between SKS and the probit reported in Sutter et al. supports the findings and main conclusions of their study, with the only difference concerning math achievement. This suggests a couple of main takeaways: first, positively that Sutter et al.'s results appear mostly robust to misspecification concerns, but second, that having supposed a probit model represents an unnecessary additional assumption in their analysis that can lead to misleading conclusions, as may be the case for math grades. We finally note that the parametric findings could be credibly supported only when validated using SKS, since the KS results were too unstable to yield reliable inference.

5 Concluding remarks

In this paper, we have proposed a simplified version of the classical KS estimator of binary choice models, labelled "SKS". We have shown that this new estimator is relatively easy to implement, is numerically more stable and performs better in simulations than the original.

When compared to parametric binary choice estimators, SKS consistently outperforms these estimators under distributional mispecification, while retaining good relative performance when the parametric estimator is correctly specified. Using the SKS estimator, we revisit the results of Sutter et al. (2013) to show that although the central finding holds, a secondary finding of that analysis is not supported by careful semiparametric analysis. We close with a brief discussion of some possible extensions and generalizations of the SKS estimator.

Endogeneity. Rothe (2009) propose a two-step semi-parametric ML estimator for binary choice models with endogenous regressors. In the first step, Rothe estimates a reduced form equation for the endogenous regressors and extract the corresponding residuals. In the second step, the residuals are added as control variables and the resulting model is estimated using the KS estimator. We conjecture that the SKS estimator can be used in place of KS in the second step. This is expected to improve small-sample properties of Rothe's estimator.

Choice of bandwidth. We have only briefly touched upon the choice of bandwidth h in the SKS estimator. Although the simulation results show that a simple fixed choice of h works well, the small-sample performance of the estimator may become even better if a data-driven approach is used. For non-parametric CDF estimation, Bowman et al. (1998) propose a cross-validation method, while Hansen (2004) proposes a refined plug-in bandwidth rule, which minimizes an estimate of the asymptotic mean integrated squared error. A third possibility – inspired by the same studies used to motivate KS2 in our simulation and empirical studies – is to treat h as an additional parameter and estimate it jointly with the coefficients of the model in the numerical optimization of the SKS objective function.

Ordered choice. Klein and Sherman (2002) generalized the results of KS to an ordered choice framework, in which the binary model considered here arise as a special case. The results in this paper can be generalized along the same lines.

References

ABO-EL-HADID, S. (2018): "Logistic kernel estimator and bandwidth selection for density function," *International Journal of Contemporary Mathematical Sciences*, 13, 279–286.

ALDRICH, J. AND F. NELSON (1984): *Linear Probability, Logit, and Probit Models*, Thousand Oaks, California: SAGE Publications, Inc.

AMEMIYA, T. (1981): "Qualitative Response Models: A Survey," *Journal of Economic Literature*, XIX, 1483–1563.

ANGRIST, J., V. CHERNOZHUKOV, AND I. FERNÁNDEZ-VAL (2006): "Quantile Regression under Misspecification, with an Application to the U.S. Wage Structure," *Econometrica*, 74, 539–563.

ANGRIST, J. D. AND J.-S. PISCHKE (2008): *Mostly harmless econometrics: An empiricist's companion*, Princeton University Press.

BANKS, J. AND Z. OLDFIELD (2007): "Understanding Pensions: Cognitive Function, Numerical Ability and Retirement Saving*," *Fiscal Studies*, 28, 143–170.

BENJAMIN, D. J., S. A. BROWN, AND J. M. SHAPIRO (2013): "Who is 'Behavioral'? Cognitive Ability and Anomalous Preferences," *Journal of the European Economic Association*, 11, 1231–1255.

BERG, A. AND D. POLITIS (2009): "CDF and survival function estimation with infinite-order kernels," *Electronic Journal of Statistics*, 3, 1436–1454.

BERNHEIM, B., D. M. GARRETT, AND D. M. MAKI (2001): "Education and saving:," *Journal of Public Economics*, 80, 435–465.

BIEWEN, M., B. FITZENBERGER, AND M. SECKLER (2020): "Counterfactual quantile decompositions with selection correction taking into account Huber/Melly (2015): An application to the German gender wage gap," *Labour Economics*, 101927.

BOWMAN, A., P. HALL, AND T. PRVAN (1998): "Bandwidth selection for the smoothing of distribution functions," *Biometrika*, 85, 799–808.

BROUNEN, D., K. G. KOEDIJK, AND R. A. POWNALL (2016): "Household financial planning and savings behavior," *Journal of International Money and Finance*, 69, 95–107.

BROWN, M., J. GRIGSBY, W. VAN DER KLAAUW, J. WEN, AND B. ZAFAR (2016): "Financial Education and the Debt Behavior of the Young," *Review of Financial Studies*, 29, 2490–2522.

CADENA, B. C. AND B. J. KEYS (2015): "Human Capital and the Lifetime Costs of Impatience," *American Economic Journal: Economic Policy*, 7, 126–153.

CAMERON, A. C. AND P. K. TRIVEDI (2005): *Microeconometrics: Methods and Applications*, New York, USA: Cambridge University Press.

CASTILLO, M., P. J. FERRARO, J. L. JORDAN, AND R. PETRIE (2011): "The today and tomorrow of kids: Time preferences and educational outcomes of children," *Journal of Public Economics*, 95, 1377–1385.

CASTILLO, M., J. L. JORDAN, AND R. PETRIE (2019): "Discount Rates of Children and High School Graduation," *The Economic Journal*, 129, 1153–1181.

CHEN, H. Z. AND A. RANDALL (1997): "Semi-parametric estimation of binary response models with an application to natural resourse valuation," *Journal of Econometrics*, 76, 323–340.

CHU, J., T.-H. LEE, AND A. ULLAH (2019): "Variable selection in sparse semiparametric single index models," *Advances in Econometrics: Topics in Identification, Limited Dependent Variables, Partial Observability, Experimentation, and Flexible Modeling*, 40b, 65–88.

COLE, S., A. PAULSON, AND G. K. SHASTRY (2016): "High School Curriculum and Financial Outcomes: The Impact of Mandated Personal Finance and Mathematics Courses," *Journal of Human Resources*, 51, 656–698.

COPPEJANS, M. (2001): "Estimation of the binary response model using a mixture of distributions estimator (MOD)," *Journal of Econometrics*, 102, 231–269.

COSSLETT, S. R. (1993): "Distribution-free maximum likelihood estimator of the binary choice model," *Econometrica*, 51, 765–782.

DELANEY, L., C. HARMON, AND M. RYAN (2013): "The role of noncognitive traits in undergraduate study behaviours," *Economics of Education Review*, 32, 181–195.

DELECROIX, M., M. HRISTACHE, AND V. PATILEA (2006): "Semiparametric *M*-estimation in single-index models," *Journal of Statistical Planning and Inference*, 136, 730–769.

DELLE SITE, P. AND J. PARMAR (2024): "On the Linear Probability Model as binary choice random utility model," *Journal of Choice Modelling*, 52, 100505.

DELUCA, G. (2008): "SNP and SML estimation of univariate and bivariate binary–choice models," *Stata Journal*, 8, 190–220.

ESTRADA-MEJIA, C., M. DE VRIES, AND M. ZEELENBERG (2016): "Numeracy and wealth," *Journal of Economic Psychology*, 54, 53–63.

FRÖLICH, M. (2006): "Nonparametric regression for binary dependent variables," *Econometrics Journal*, 9, 511–540.

GERFIN, M. (1996): "Parametric and semi-parametric estimation of the binary response model of labour market participation," *Journal of Applied Econometrics*, 11, 321–339.

GOLSTEYN, B. H., H. GRÖNQVIST, AND L. LINDAHL (2014): "Adolescent Time Preferences Predict Lifetime Outcomes," *The Economic Journal*, 124, F739–F761.

HANSEN, B. E. (2004): "Nonparametric estimation of smooth conditional distributions," .

——— (2008): "Uniform convergence rates for kernel estimation with dependent data," *Econometric Theory*, 24, 726–748.

HANUSHEK, E. A., L. KINNE, P. LERGETPORER, AND L. WOESSMANN (2021): "Patience, Risk-Taking, and Human Capital Investment Across Countries," *The Economic Journal*, 132, 2290–2307.

HANUSHEK, E. A., L. KINNE, P. SANCASSANI, AND L. WOESSMANN (2023): "Can Patience Account for Subnational Differences in Student Achievement? Regional Analysis with Facebook Interests," Working Paper 31690, National Bureau of Economic Research.

HÄRDLE, W., P. HALL, AND H. ICHMURA (1993): "Optimal smoothing in single-index models," *Annals of Statistics*, 21, 157–178.

HORN, D. AND H. J. KISS (2020): "Time preferences and their life outcome correlates: Evidence from a representative survey," *PLOS ONE*, 15, e0236486.

HOROWITZ, J. AND W. HÄRDLE (1996): "Direct semiparametric estimation of single-index models with discrete covariates," *Journal of the American Statistical Association*, 91, 1632–1640.

ICHIMURA, H. (1993): "Semiparametric least squares estimation of single index models (SLS) and weighted SLS estimation of single index models," *Journal of Econometrics*, 58, 71–120.

ICHIMURA, H. AND P. TODD (2007): "Implementing nonparametric and semiparametric estimators," in *Handbook of Econometrics*, Elsevier, vol. 6, 5369–5468.

KLEIN, R. AND R. P. SHERMAN (2002): "Shift restrictions and semiparametric estimation in ordered response models," *Econometrica*, 70, 663–691.

KLEIN, R. AND R. SPADY (1993): "An efficient semiparametric estimator for binary response models," *Econometrica*, 61, 387–421.

LEE, L. (1995): "Semiprametric maximum likelihood estimation of polychotomous and sequential choice models," *Journal of Econometrics*, 65, 381–428.

LEWBEL, A., Y. DONG, AND T. T. YANG (2012): "Comparing features of convenient estimators for binary choice models with endogenous regressors," *Canadian Journal of Economics*, 45, 809–829.

LI, Q. AND J. S. RACINE (2007): *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

LUSARDI, A. (2012): *Numeracy, financial literacy, and financial decision-making*, w17821, Cambridge, MA.

LÜHRMANN, M., M. SERRA-GARCIA, AND J. WINTER (2018): "The Impact of Financial Education on Adolescents' Intertemporal Choices," *American Economic Journal: Economic Policy*, 10, 309–332.

MARLEY-PAYNE, J., P. DITURI, AND A. DAVIDSON (2022): "Financial Education, Mathematical Confidence, and Financial Behavior," *Journal of Financial Counseling and Planning*, 33, 194–204.

MENEZES-FILHO, N. A., M.-A. MUENDLER, AND G. RAMEY (2008): "The structure of worker compensation in Brazil, with an comparison to France and the United States," *The Review of Economics and Statistics*, 90, 324–346.

MITTELHAMMER, R. C. AND G. JUDGE (2011): "A family of empirical likelihood functions and estimators for the binary response model," *Journal of Econometrics*, 164, 207–217.

MOFFITT, R. A. (1999): "Chapter 24 - New Developments in Econometric Methods for Labor Market Analysis," Elsevier, vol. 3 of *Handbook of Labor Economics*, 1367–1397.

NEWEY, W. K. AND D. MCFADDEN (1994): "Large sample estimation and hypothesis testing," in *Handbook of Econometrics*, Elsevier, vol. 6, 2112–2245.

NEWEY, W. K., J. L. POWELL, AND J. R. WALKER (1990): "Semiparametric estimation of selection models: Some empirical results," *The American Economic Review, Papers and Proceedings*, 80, 324–328.

ROTHE, C. (2009): "Semiparametric estimation of binary response models with endogenous regressors," *Journal of Econometrics*, 153, 51–64.

SUTTER, M., M. KOCHER, D. GLÄTZLE-RÜTZLER, AND S. TRAUTMANN (2013): "Impatience and uncertainty: Experimental decisions predict adolescents' field behavior," *American Economic Review*, 103, 510–531.

SUTTER, M., C. ZOLLER, AND D. GLÄTZLE-RÜTZLER (2019): "Economic behavior of children and adolescents – A first survey of experimental economics results," *European Economic Review*, 111, 98–121.

WESTERLUND, J. AND P. HJERTSTRAND (2014): "Indirect estimation of semiparametric binary choice models," *Oxford Bulletin of Economics and Statistics*, 76, 298–314.

WINTER, B. B. (1979): "Convergence rate of perturbed empirical distribution functions," *Journal of Applied Probability*, 16, 163–173.

A Proofs

Our asymptotic results follow closely those of KS. We begin by deriving uniform consistency results for $\hat{F}(x)$ and its two first derivatives. This is Lemma A.1, which is our version of Lemma 2 in KS.

Lemma A.1 (Kernel consistency)**.** *Suppose that Assumptions 1–3, 6 and 7 are met. Then,*

$$\sup_{v\in\mathbb{R}} |\hat{F}(v) - F(v)| = O_p\left(\sqrt{\frac{\ln\ln n}{n}}\right), \tag{A.1}$$

$$\sup_{v\in\mathbb{R}} \left|\frac{d\hat{F}(v)}{dv} - \frac{dF(v)}{dv}\right| = O_p\left(\sqrt{\frac{\ln n}{nh}} + h^2\right), \tag{A.2}$$

$$\sup_{v\in\mathbb{R}} \left|\frac{d^2\hat{F}(v)}{(dv)^2} - \frac{d^2F(v)}{(dv)^2}\right| = O_p\left(\sqrt{\frac{\ln n}{nh^3}} + h^2\right). \tag{A.3}$$

Proof. Consider $\hat{F}(v)$. We have $\int v^2 k(v)dv < \infty$ and since $k(v) = k(-v)$ (symmetry) we also have $\int vk(v)dv = 0$. Hence, if we in addition assume that h is such that $h^2\sqrt{n/\ln\ln n} \to 0$, then by Theorem 3.2 of Winter (1979), $\hat{F}(v)$ has the so-called "Chung–Smirnov property". This means that

$$\sup_{v\in\mathbb{R}} |\hat{F}(v) - F(v)| = O_p\left(\sqrt{\frac{\ln\ln n}{n}}\right). \tag{A.4}$$

In order to see what $h^2\sqrt{n/\ln\ln n} \to 0$ means for h, note that $h^2\sqrt{n/\ln\ln n} \geq 0$. Hence, requiring that $h^2\sqrt{n/\ln\ln n} \to 0$ is equivalent to requiring $h^4 n/\ln\ln n \to 0$, which will be the case if $h \sim n^{-\alpha}$ with $\alpha \geq 1/4$.

We now turn to the derivatives. Clearly, since $\hat{F}(x)$ is an estimator of the CDF $F(x)$, $d\hat{F}(x)/dx = \hat{f}(x)$ is an estimator of the PDF $f(v)$. This means that the rates of consistency of $d\hat{F}(v)/dv$ and $d^2\hat{F}(v)/(dv)^2$ can be taken from Hansen (2008), who establishes uniform rates for kernel estimators of PDFs when data are weakly dependent. In particular, by his Theorem

6 for regular second-order kernels and scalar v,

$$\sup_{v\in\mathbb{R}}\left|\frac{d\hat{F}(v)}{dv}-\frac{dF(v)}{dv}\right|=\sup_{v\in\mathbb{R}}|\hat{f}(v)-f(v)|=O_p\left(\sqrt{\frac{\ln n}{nh}}+h^2\right), \tag{A.5}$$

$$\sup_{v\in\mathbb{R}}\left|\frac{d^2\hat{F}(v)}{(dv)^2}-\frac{d^2F(v)}{(dv)^2}\right|=\sup_{v\in\mathbb{R}}\left|\frac{d\hat{f}(v)}{dv}-\frac{df(v)}{dv}\right|=O_p\left(\sqrt{\frac{\ln n}{nh^3}}+h^2\right), \tag{A.6}$$

which holds provided only that $h=o(1)$. These results therefore hold under our condition that $h=O(n^{-1/4})$. This establishes the required rates for the derivatives and hence the proof of the lemma is complete. □

Proof of Theorem 1. To establish consistency, we take the usual route (see, for example, KS, Proof of Theorem 3) and first show that the estimated and normalized likelihood function $n^{-1}\hat{\ell}_{\text{SKS}}(\beta)$ converges uniformly to $n^{-1}\ell(\beta)$. We then show that $n^{-1}\ell(\beta)$ attains a unique maximum at $\beta=\beta^0$, which implies both that β^0 is identified and that $\hat{\beta}$ is consistent.

We begin by noting that since n does not depend on β, scaling of the objective function by this quantity is inconsequential. We therefore proceed to evaluate $\hat{\ell}_{\text{SKS}}(\beta)/n$;

$$n^{-1}\hat{\ell}_{\text{SKS}}(\beta)=\frac{1}{n}\sum_{i=1}^{n}[y_i\ln\hat{F}(x_i'\beta)+(1-y_i)\ln(1-\hat{F}(x_i'\beta))]. \tag{A.7}$$

Consider $n^{-1}\sum_{i=1}^{n}y_i\ln\hat{F}(x_i'\beta)$. By Taylor expanding $\ln\hat{F}(v)$ about $F(v)$,

$$\ln\hat{F}(v)=\ln F(v)+F(v)^{-1}[\hat{F}(v)-F(v)]+O_p([\hat{F}(v)-F(v)]^2). \tag{A.8}$$

Further use of Lemma A.1 and the assumption that $F(v)>0$ uniformly in v yields

$$\sup_{v\in\mathbb{R}}|\ln\hat{F}(v)-\ln F(v)|=O_p(\hat{F}(v)-F(v))=O_p\left(\sqrt{\frac{\ln\ln n}{n}}\right)=o_p(1), \tag{A.9}$$

which holds provided $h\sim n^{-\alpha}$ with $\alpha\geq 1/4$. By using this, the triangle inequality and

$y_i \in \{0,1\}$,

$$\begin{aligned}
&\sup_\beta \left| \frac{1}{n}\sum_{i=1}^n y_i \ln \hat{F}(x_i'\beta) - \frac{1}{n}\sum_{i=1}^n y_i \ln F(x_i'\beta) \right| \\
&= \sup_\beta \left| \frac{1}{n}\sum_{i=1}^n y_i [\ln \hat{F}(x_i'\beta) - \ln F(x_i'\beta)] \right| \\
&\leq \frac{1}{n}\sum_{i=1}^n y_i \sup_\beta |\ln \hat{F}(x_i'\beta) - \ln F(x_i'\beta)| \\
&\leq \frac{1}{n}\sum_{i=1}^n \sup_\beta |\ln \hat{F}(x_i'\beta) - \ln F(x_i'\beta)| = o_p(1). \qquad \text{(A.10)}
\end{aligned}$$

The same arguments can be used to show that

$$\sup_\beta \left| \frac{1}{n}\sum_{i=1}^n (1-y_i) \ln(1-\hat{F}(x_i'\beta)) - \frac{1}{n}\sum_{i=1}^n (1-y_i) \ln(1-F(x_i'\beta)) \right| = o_p(1). \qquad \text{(A.11)}$$

Hence, by adding the results,

$$\sup_\beta |n^{-1}\hat{\ell}_{\text{SKS}}(\beta) - n^{-1}\ell(\beta)| = o_p(1). \qquad \text{(A.12)}$$

The rest of the proof follows by the same arguments used by Rothe (2009, Proof of Theorem 2). Note in particular that since $\hat{\ell}(\beta)$ is an ordinary parametric log-likelihood function, by a standard uniform law of large numbers (see, for example, Newey and McFadden 1994, Lemma 2.4), it converges uniformly in β to its expectation;

$$\sup_\beta |n^{-1}\ell(\beta) - n^{-1}\mathbb{E}[\ell(\beta)]| = o_p(1), \qquad \text{(A.13)}$$

with

$$\begin{aligned}
n^{-1}\mathbb{E}[\ell(\beta)] &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[y_i \ln F(x_i'\beta) + (1-y_i)\ln(1-F(x_i'\beta))] \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\mathbb{E}(y_i|x_i) \ln F(x_i'\beta) + (1-\mathbb{E}(y_i|x_i))\ln(1-F(x_i'\beta))] \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[F(x_i'\beta^0) \ln F(x_i'\beta) + (1-F(x_i'\beta^0))\ln(1-F(x_i'\beta))] \\
&= \mathbb{E}[F(x_1'\beta^0) \ln F(x_1'\beta) + (1-F(x_1'\beta^0))\ln(1-F(x_1'\beta))], \qquad \text{(A.14)}
\end{aligned}$$

where the last equality holds because x_i is iid over i. The last expectation attains its maximum whenever $F(x_1'\beta^0) = F(x_1'\beta)$, which by assumption can only happen if $\beta = \beta^0$. Consistency now follows from standard arguments (see, for example, Newey and McFadden 1994, Theorem 2.1). □

Proof of Lemma 1. By using Lemma A.1, the fact that $\|\beta^* - \beta^0\| \leq \|\hat{\beta} - \beta^0\| = o_p(1)$ by Theorem 1 and the steps used in the proof of that theorem to establish the uniform convergence of $n^{-1}\hat{\ell}_{\text{SKS}}(\beta)$ to $n^{-1}\hat{\ell}(\beta)$, we can show that

$$\sup_{\beta} \left\| \frac{1}{n} \frac{d^2\hat{\ell}_{\text{SKS}}(\beta^*)}{d\beta(d\beta)'} - \frac{1}{n} \frac{d^2\ell(\beta^0)}{d\beta(d\beta)'} \right\| = O_p\left(\sqrt{\frac{\ln n}{nh^3}} + h^2 \right). \tag{A.15}$$

By using l'Hôpital's rule, we can show that $\ln n/(nh^3) + h^2 \to 0$ if $h \sim n^{-\alpha}$ with $0 < \alpha < 1/3$, which is clearly the case under our assumption about the rate of shrinking of h. Hence, $O_p(\sqrt{\ln n/(nh^3)} + h^2) = o_p(1)$.

Let us now consider $n^{-1}d^2\ell(\beta^0)/[d\beta(d\beta)']$. A direct calculation yields

$$\frac{d\ell(\beta)}{d\beta} = \sum_{i=1}^{n} \frac{y_i - F(x_i'\beta)}{F(x_i'\beta)[1 - F(x_i'\beta)]} \frac{dF(x_i'\beta)}{d\beta}, \tag{A.16}$$

which can be differentiated again to obtain

$$\begin{aligned} \frac{d^2\ell(\beta)}{d\beta(d\beta)'} = \sum_{i=1}^{n} \Bigg[& - \frac{1}{F(x_i'\beta)[1 - F(x_i'\beta)]} \frac{dF(x_i'\beta)}{d\beta} \left(\frac{dF(x_i'\beta)}{d\beta} \right)' \\ & - \frac{[1 - 2F(x_i'\beta)][y_i - F(x_i'\beta)]}{F(x_i'\beta)^2[1 - F(x_i'\beta)]^2} \frac{dF(x_i'\beta)}{d\beta} \left(\frac{dF(x_i'\beta)}{d\beta} \right)' \\ & + \frac{y_i - F(x_i'\beta)}{F(x_i'\beta)[1 - F(x_i'\beta)]} \frac{d^2F(x_i'\beta)}{d\beta(d\beta)'} \Bigg]. \end{aligned} \tag{A.17}$$

Hence, by a law of large numbers for iid variates,

$$\begin{aligned}
\frac{1}{n}\frac{d^2\ell(\beta)}{d\beta(d\beta)'} \to_p \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Bigg[& -\frac{1}{F(x_i'\beta)[1-F(x_i'\beta)]}\frac{dF(x_i'\beta)}{d\beta}\left(\frac{dF(x_i'\beta)}{d\beta}\right)' \\
& -\frac{[1-2F(x_i'\beta)][y_i-F(x_i'\beta)]}{F(x_i'\beta)^2[1-F(x_i'\beta)]^2}\frac{dF(x_i'\beta)}{d\beta}\left(\frac{dF(x_i'\beta)}{d\beta}\right)' \\
& +\frac{y_i-F(x_i'\beta)}{F(x_i'\beta)[1-F(x_i'\beta)]}\frac{d^2F(x_i'\beta)}{d\beta(d\beta)'}\Bigg] \\
= \mathbb{E}\Bigg[& -\frac{1}{F(x_1'\beta)[1-F(x_1'\beta)]}\frac{dF(x_1'\beta)}{d\beta}\left(\frac{dF(x_1'\beta)}{d\beta}\right)' \\
& -\frac{[1-2F(x_1'\beta)][y_1-F(x_1'\beta)]}{F(x_1'\beta)^2[1-F(x_1'\beta)]^2}\frac{dF(x_1'\beta)}{d\beta}\left(\frac{dF(x_1'\beta)}{d\beta}\right)' \\
& +\frac{y_1-F(x_1'\beta)}{F(x_1'\beta)[1-F(x_1'\beta)]}\frac{d^2F(x_1'\beta)}{d\beta(d\beta)'}\Bigg]
\end{aligned} \tag{A.18}$$

as $n \to \infty$. Let $u_i := y_i - F(x_i'\beta^0) = y_i - \mathbb{E}(y_i|x_i)$, which under our conditions is iid with $\mathbb{E}(u_i|x_i) = 0$ and $\text{var}(u_i|x_i) = \text{var}(y_i|x_i) = F(x_i'\beta^0)[1-F(x_i'\beta^0)]$ (see, for example, Ichimura 1993). By using this and the law of iterated expectations, the above limiting expression for $n^{-1}d^2\ell(\beta^0)/[d\beta(d\beta)']$ at $\beta = \beta^0$ reduces to

$$\begin{aligned}
\frac{1}{n}\frac{d^2\ell(\beta^0)}{d\beta(d\beta)'} \to_p \mathbb{E}\Bigg[& -\frac{1}{F(x_1'\beta^0)[1-F(x_1'\beta^0)]}\frac{dF(x_1'\beta^0)}{d\beta}\left(\frac{dF(x_1'\beta^0)}{d\beta}\right)' \\
& -\frac{[1-2F(x_1'\beta^0)]u_1}{F(x_1'\beta^0)^2[1-F(x_1'\beta^0)]^2}\frac{dF(x_1'\beta^0)}{d\beta}\left(\frac{dF(x_1'\beta^0)}{d\beta}\right)' \\
& +\frac{u_1}{F(x_1'\beta^0)[1-F(x_1'\beta^0)]}\frac{d^2F(x_1'\beta^0)}{d\beta(d\beta)'}\Bigg] \\
= \mathbb{E}\Bigg[& -\frac{1}{F(x_1'\beta^0)[1-F(x_1'\beta^0)]}\frac{dF(x_1'\beta^0)}{d\beta}\left(\frac{dF(x_1'\beta^0)}{d\beta}\right)' \\
& -\frac{[1-2F(x_1'\beta^0)]\mathbb{E}(u_1|x_1)}{F(x_1'\beta^0)^2[1-F(x_1'\beta^0)]^2}\frac{dF(x_1'\beta^0)}{d\beta}\left(\frac{dF(x_1'\beta^0)}{d\beta}\right)' \\
& +\frac{\mathbb{E}(u_1|x_1)}{F(x_1'\beta^0)[1-F(x_1'\beta^0)]}\frac{d^2F(x_1'\beta^0)}{d\beta(d\beta)'}\Bigg] \\
= -\mathbb{E}\Bigg[& \frac{1}{F(x_1'\beta^0)[1-F(x_1'\beta^0)]}\frac{dF(x_1'\beta^0)}{d\beta}\left(\frac{dF(x_1'\beta^0)}{d\beta}\right)'\Bigg].
\end{aligned} \tag{A.19}$$

By adding the results,

$$\begin{aligned}\frac{1}{n}\frac{d^2\hat{\ell}_{\text{SKS}}(\beta^*)}{d\beta(d\beta)'} &= \frac{1}{n}\frac{d^2\ell(\beta^0)}{d\beta(d\beta)'} + o_p(1) \\ &\to_p -\mathbb{E}\left[\frac{1}{F(x_1'\beta^0)[1-F(x_1'\beta^0)]}\frac{dF(x_1'\beta^0)}{d\beta}\left(\frac{dF(x_1'\beta^0)}{d\beta}\right)'\right] \end{aligned} \tag{A.20}$$

as $n \to \infty$, which is what we wanted to show. □

Proof of Lemma 2. This proof follows closely the proof of Theorem 5.2 in Ichimura (1993). From the expression for $d\ell(\beta^0)/d\beta$ given in Proof of Lemma 1,

$$\begin{aligned}&\frac{d\hat{\ell}_{\text{SKS}}(\beta^0)}{d\beta} - \frac{d\ell(\beta^0)}{d\beta} \\ &= \sum_{i=1}^n \frac{\hat{u}_i}{\hat{F}(x_i'\beta^0)[1-\hat{F}(x_i'\beta^0)]}\frac{d\hat{F}(x_i'\beta^0)}{d\beta} - \sum_{i=1}^n \frac{u_i}{F(x_i'\beta^0)[1-F(x_i'\beta^0)]}\frac{dF(x_i'\beta^0)}{d\beta} \\ &= \sum_{i=1}^n \left(\frac{\hat{u}_i}{\hat{F}(x_i'\beta^0)[1-\hat{F}(x_i'\beta^0)]} - \frac{u_i}{F(x_i'\beta^0)[1-F(x_i'\beta^0)]}\right)\frac{d\hat{F}(x_i'\beta^0)}{d\beta} \\ &+ \sum_{i=1}^n \frac{u_i}{F(x_i'\beta^0)[1-F(x_i'\beta^0)]}\left(\frac{d\hat{F}(x_i'\beta^0)}{d\beta} - \frac{dF(x_i'\beta^0)}{d\beta}\right), \end{aligned} \tag{A.21}$$

where $\hat{u}_i := y_i - \hat{F}(x_i'\beta^0)$ and $u_i := y_i - F(x_i'\beta^0)$ as in Proof of Lemma 2. Consider the second term on the right-hand side, which as the same form as the one in Lemma 5.8 of Ichimura (1993). We therefore use the steps as in the proof of that lemma to show that the second term above is $o_p(1)$ when divided by $\sqrt{n}$. We begin by noting that the mean is zero. As for the variance, making use of the definition of $\hat{F}(x_i'\beta)$ as $\hat{F}(x_i'\beta) := (n-1)^{-1}\sum_{j\neq i}^n K_h((x_j - x_i)'\beta)$ where $K_h(v) := K(v)/h$, we can write

$$\frac{u_i}{F(x_i'\beta^0)[1-F(x_i'\beta^0)]}\left(\frac{d\hat{F}(x_i'\beta^0)}{d\beta} - \frac{dF(x_i'\beta^0)}{d\beta}\right) = \frac{1}{n-1}\sum_{j\neq i}^n u_i a_{ij} \tag{A.22}$$

where

$$\begin{aligned} a_{ij} &:= \frac{1}{F(x_i'\beta^0)[1-F(x_i'\beta^0)]}\left(\frac{dK_h((x_j-x_i)'\beta^0)}{d\beta} - \frac{dF(x_i'\beta^0)}{d\beta}\right) \\ &= \frac{1}{F(x_i'\beta^0)[1-F(x_i'\beta^0)]}\left(\frac{k_h((x_j-x_i)'\beta^0)(x_j-x_i)}{h} - \frac{dF(x_i'\beta^0)}{d\beta}\right) \end{aligned} \tag{A.23}$$

with $k_h(v) = dK_h(v)/dv$. In this notation, the second term on the right-hand side of (A.21) divided by $\sqrt{n}$ is

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{u_i}{F(x_i'\beta^0)[1-F(x_i'\beta^0)]}\left(\frac{d\hat{F}(x_i'\beta^0)}{d\beta}-\frac{dF(x_i'\beta^0)}{d\beta}\right)=\frac{1}{\sqrt{n}(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}u_i a_{ij}. \quad \text{(A.24)}$$

Making use of this and the fact that u_i is iid, the sought variance is given by

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{\sqrt{n}(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}u_i a_{ij}\left(\frac{1}{\sqrt{n}(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}u_i a_{ij}\right)'\right]\\
&=\frac{1}{n(n-1)^2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\sum_{k=1}^{n}\sum_{l\neq k}^{n}\mathbb{E}(u_i u_k a_{ij}a_{kl}')\\
&=\frac{n-2}{n-1}\mathbb{E}(u_i^2 a_{ij}a_{ik}')+\frac{1}{n-1}\mathbb{E}(u_i^2 a_{ij}a_{ij}')+\frac{1}{n-1}\mathbb{E}(u_i u_j a_{ij}a_{ji}'), \quad \text{(A.25)}
\end{aligned}$$

where i, j and k are different (see Ichimura 1993, page 116). The first term on the right dominate the other two. We therefore consider this term first. Because i, j and k are different,

$$\begin{aligned}
&\mathbb{E}(u_i^2 a_{ij}a_{ik}')\\
&=\mathbb{E}\left\{\frac{u_i^2}{F(x_i'\beta^0)^2[1-F(x_i'\beta^0)]^2}\mathbb{E}\left[\left(\frac{k_h((x_j-x_i)'\beta^0)(x_j-x_i)}{h}-\frac{dF(x_i'\beta^0)}{d\beta}\right)\right.\right.\\
&\left.\left.\times\left(\frac{k_h((x_k-x_i)'\beta^0)(x_k-x_i)}{h}-\frac{dF(x_i'\beta^0)}{d\beta}\right)'|y_i,x_i\right]\right\}\\
&=\mathbb{E}\left\{\frac{u_i^2}{F(x_i'\beta^0)^2[1-F(x_i'\beta^0)]^2}\left(\mathbb{E}\left[\frac{k_h((x_j-x_i)'\beta^0)(x_j-x_i)}{h}|y_i,x_i\right]-\frac{dF(x_i'\beta^0)}{d\beta}\right)\right.\\
&\left.\times\left(\mathbb{E}\left[\frac{k_h((x_k-x_i)'\beta^0)(x_k-x_i)}{h}|y_i,x_i\right]-\frac{dF(x_i'\beta^0)}{d\beta}\right)'\right\}, \quad \text{(A.26)}
\end{aligned}$$

where the bracketed terms are $O_p(h^2)$ by Lemma A.2 of Ichimura (1993). It follows that

$$\frac{n-2}{n-1}\mathbb{E}(u_i^2 a_{ij}a_{ik}')=O(h^4)=o(1), \quad \text{(A.27)}$$

since $h=o(1)$. The last terms on the right-hand side of (A.25) are $o(1)$ too, provided $nh^2\to\infty$.

This is so because $ha_{ij} = O_p(1)$ and hence

$$\frac{1}{n-1}\mathbb{E}(u_i^2 a_{ij} a_{ij}') = \frac{1}{(n-1)h^2}\mathbb{E}(u_i^2 h a_{ij} h a_{ij}') = o_p(1) \tag{A.28}$$

if $nh^2 \to \infty$. Therefore, since the variance is $o(1)$, we have

$$\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{u_i}{F(x_i'\beta^0)[1 - F(x_i'\beta^0)]} \left(\frac{d\hat{F}(x_i'\beta^0)}{d\beta} - \frac{dF(x_i'\beta^0)}{d\beta} \right) \right\| = o_p(1) \tag{A.29}$$

We now continue onto the first term on the right-hand side of (A.21). From

$$\begin{aligned}
&\hat{F}(x_i'\beta)[1 - \hat{F}(x_i'\beta)]F(x_i'\beta)[1 - F(x_i'\beta)] \left(\frac{\hat{u}_i}{\hat{F}(x_i'\beta)[1 - \hat{F}(x_i'\beta)]} - \frac{u_i}{F(x_i'\beta)[1 - F(x_i'\beta)]} \right) \\
&= \hat{u}_i F(x_i'\beta)[1 - F(x_i'\beta)] - u_i \hat{F}(x_i'\beta)[1 - \hat{F}(x_i'\beta)] \\
&= -[\hat{F}(x_i'\beta) - F(x_i'\beta)]F(x_i'\beta)[1 - F(x_i'\beta)] \\
&+ u_i(\hat{F}(x_i'\beta)[1 - \hat{F}(x_i'\beta)] - F(x_i'\beta)[1 - F(x_i'\beta)]) \\
&= (\hat{u}_i - u_i)F(x_i'\beta)[1 - F(x_i'\beta)] + u_i([\hat{F}(x_i'\beta) - F(x_i'\beta)][1 - \hat{F}(x_i'\beta)] \\
&- F(x_i'\beta)[\hat{F}(x_i'\beta) - F(x_i'\beta)]) \\
&= -[\hat{F}(x_i'\beta) - F(x_i'\beta)]F(x_i'\beta)[1 - F(x_i'\beta)] \\
&+ u_i[\hat{F}(x_i'\beta) - F(x_i'\beta)][1 - \hat{F}(x_i'\beta) - F(x_i'\beta)],
\end{aligned} \tag{A.30}$$

we obtain

$$
\begin{aligned}
&\sum_{i=1}^{n}\left(\frac{\hat{u}_i}{\hat{F}(x_i'\beta^0)[1-\hat{F}(x_i'\beta^0)]}-\frac{u_i}{F(x_i'\beta^0)[1-F(x_i'\beta^0)]}\right)\frac{d\hat{F}(x_i'\beta^0)}{d\beta}\\
&=-\sum_{i=1}^{n}\frac{[\hat{F}(x_i'\beta)-F(x_i'\beta)]}{\hat{F}(x_i'\beta)[1-\hat{F}(x_i'\beta)]}\frac{d\hat{F}(x_i'\beta^0)}{d\beta}\\
&+\sum_{i=1}^{n}\frac{u_i[\hat{F}(x_i'\beta)-F(x_i'\beta)][1-\hat{F}(x_i'\beta)-F(x_i'\beta)]}{\hat{F}(x_i'\beta)[1-\hat{F}(x_i'\beta)]F(x_i'\beta)[1-F(x_i'\beta)]}\frac{d\hat{F}(x_i'\beta^0)}{d\beta}\\
&=-\sum_{i=1}^{n}\frac{[\hat{F}(x_i'\beta)-F(x_i'\beta)]}{\hat{F}(x_i'\beta)[1-\hat{F}(x_i'\beta)]}\frac{dF(x_i'\beta^0)}{d\beta}\\
&+\sum_{i=1}^{n}\frac{u_i[\hat{F}(x_i'\beta)-F(x_i'\beta)][1-\hat{F}(x_i'\beta)-F(x_i'\beta)]}{\hat{F}(x_i'\beta)[1-\hat{F}(x_i'\beta)]F(x_i'\beta)[1-F(x_i'\beta)]}\frac{dF(x_i'\beta^0)}{d\beta}\\
&-\sum_{i=1}^{n}\frac{[\hat{F}(x_i'\beta)-F(x_i'\beta)]}{\hat{F}(x_i'\beta)[1-\hat{F}(x_i'\beta)]}\left(\frac{d\hat{F}(x_i'\beta^0)}{d\beta}-\frac{dF(x_i'\beta^0)}{d\beta}\right)\\
&+\sum_{i=1}^{n}\frac{u_i[\hat{F}(x_i'\beta)-F(x_i'\beta)][1-\hat{F}(x_i'\beta)-F(x_i'\beta)]}{\hat{F}(x_i'\beta)[1-\hat{F}(x_i'\beta)]F(x_i'\beta)[1-F(x_i'\beta)]}\left(\frac{d\hat{F}(x_i'\beta^0)}{d\beta}-\frac{dF(x_i'\beta^0)}{d\beta}\right). \qquad \text{(A.31)}
\end{aligned}
$$

The terms that appear here are analogous to those considered in Lemmas 5.8–5.10 of Ichimura (1993). They are $o_p(1)$ when divided by $\sqrt{n}$. Together with the triangle inequality, the above results imply

$$
\left\|\frac{1}{\sqrt{n}}\frac{d\hat{\ell}_{\mathrm{SKS}}(\beta^0)}{d\beta}-\frac{1}{\sqrt{n}}\frac{d\ell(\beta^0)}{d\beta}\right\|=o_p(1). \qquad \text{(A.32)}
$$

Moreover, by standard arguments (see, for example, Ichimura 1993, Proof of Theorem 5.2, or KS, Proof of Theorem 4),

$$
\frac{1}{\sqrt{n}}\frac{d\ell(\beta^0)}{d\beta}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{u_i}{F(x_i'\beta^0)[1-F(x_i'\beta^0)]}\frac{dF(x_i'\beta^0)}{d\beta}\to_d N(0_{r\times 1},\Sigma) \qquad \text{(A.33)}
$$

as $n \to \infty$, where

$$\begin{aligned}
\Sigma &:= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{E}\left[\frac{\mathbb{E}(u_i u_j | x_i, x_j)}{F(x_i'\beta^0)^2[1 - F(x_i'\beta^0)] F(x_j'\beta^0)^2[1 - F(x_j'\beta^0)]} \right. \\
&\quad \left. \times \frac{dF(x_i'\beta^0)}{d\beta} \left(\frac{dF(x_j'\beta^0)}{d\beta} \right)' \right] \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\frac{\mathbb{E}(u_i^2 | x_i)}{F(x_i'\beta^0)^2[1 - F(x_i'\beta^0)]^2} \frac{dF(x_i'\beta^0)}{d\beta} \left(\frac{dF(x_i'\beta^0)}{d\beta} \right)' \right] \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\frac{1}{F(x_i'\beta^0)[1 - F(x_i'\beta^0)]} \frac{dF(x_i'\beta^0)}{d\beta} \left(\frac{dF(x_i'\beta^0)}{d\beta} \right)' \right] \\
&= \mathbb{E}\left[\frac{1}{F(x_1'\beta^0)[1 - F(x_1'\beta^0)]} \frac{dF(x_1'\beta^0)}{d\beta} \left(\frac{dF(x_1'\beta^0)}{d\beta} \right)' \right], \qquad \text{(A.34)}
\end{aligned}$$

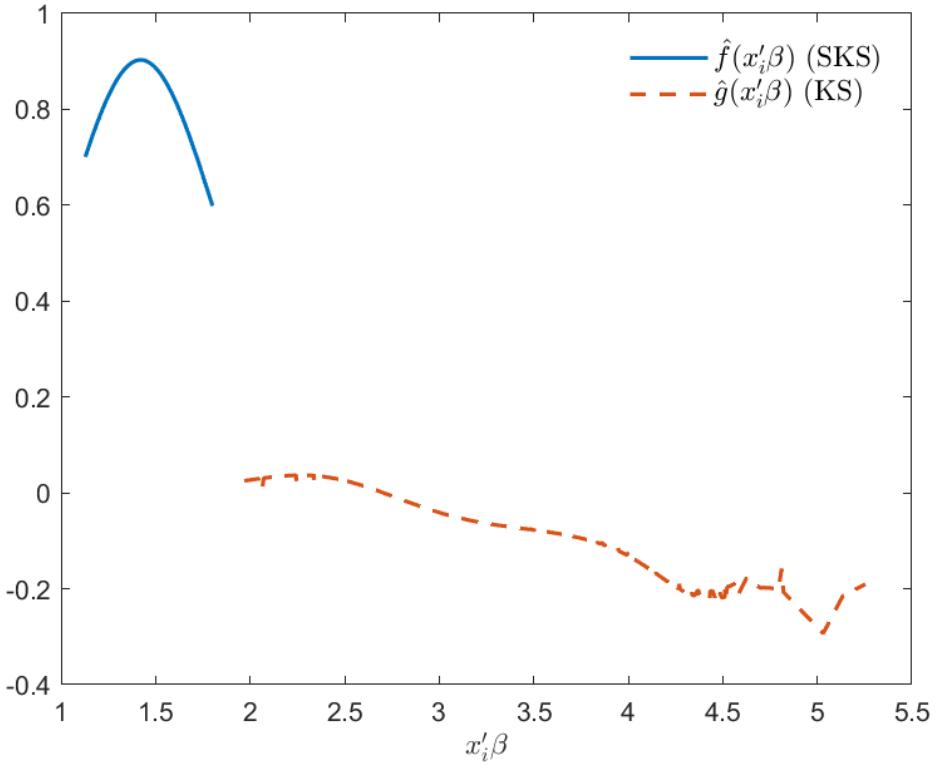
which holds since $\mathbb{E}(u_i u_j | x_i, x_j) = 0$ for $i \neq j$ and $\mathbb{E}(u_i^2 | x_i) = F(x_i'\beta^0)[1 - F(x_i'\beta^0)]$ (see Proof of Lemma 1). The fourth equality is due to iid-ness.

The asymptotic normality of $n^{-1/2} d\ell(\beta^0)/d\beta$ and the convergence of $n^{-1/2} d\hat{\ell}_{\text{SKS}}(\beta^0)/d\beta$ to $n^{-1/2} d\ell(\beta^0)/d\beta$ imply

$$\frac{1}{\sqrt{n}} \frac{d\hat{\ell}_{\text{SKS}}(\beta^0)}{d\beta} = \frac{1}{\sqrt{n}} \frac{d\ell(\beta^0)}{d\beta} + o_p(1) \to_d N(0_{r\times 1}, \Sigma), \qquad \text{(A.35)}$$

as required. □

Figure 1: Plotting $\hat{g}(x_i'\beta)$ and $\hat{f}(x_i'\beta)$ over $x_i'\beta$.



Notes: While the horizontal axis represents values of $x_i'\beta$, the vertical axis represents values of $\hat{g}(x_i'\beta)$ and $\hat{f}(x_i'\beta)$. The blue (solid) and red (dashed) lines are for $\hat{f}(x_i'\beta)$ and $\hat{g}(x_i'\beta)$, respectively.

Table 1: Aggregated Monte Carlo results for E1.

		Bias			RMSE			MAD		
Start	Estimator	$n = 250$	$n = 500$	$n = 1500$	$n = 250$	$n = 500$	$n = 1500$	$n = 250$	$n = 500$	$n = 1500$
	LPM	−0.113	−0.109	−0.108	0.955	0.752	0.677	0.357	0.313	0.300
S1	probit	−0.001	−0.001	0.000	0.333	0.216	0.120	0.190	0.132	0.074
	KS1	−0.009	−0.007	−0.003	0.322	0.206	0.110	0.125	0.084	0.045
	KS2	−0.002	−0.001	0.000	0.370	0.261	0.153	0.202	0.150	0.091
	SKS	0.004	0.002	0.001	0.348	0.224	0.122	0.200	0.136	0.074
S2	probit	−0.041	0.014	0.021	32.325	0.552	0.423	0.246	0.189	0.126
	KS1	−0.067	−0.060	−0.060	0.808	0.545	0.504	0.274	0.190	0.112
	KS2	−0.026	−0.028	−0.034	0.640	0.426	0.381	0.236	0.166	0.097
	SKS	0.003	0.002	0.001	0.464	0.229	0.125	0.205	0.139	0.077
S3	KS1	−0.062	−0.008	−0.003	32.324	0.428	0.175	0.218	0.152	0.087
	KS2	−0.058	−0.003	0.000	32.322	0.312	0.164	0.233	0.163	0.095
	SKS	−0.039	0.001	0.001	32.413	0.379	0.125	0.205	0.139	0.077
S4	probit	0.012	0.022	0.010	1.282	2.430	0.797	0.346	0.282	0.225
	KS1	0.623	0.596	0.500	1.443	1.287	1.081	0.584	0.427	0.240
	KS2	0.205	0.239	0.216	11.152	15.805	0.912	0.373	0.240	0.124
	SKS	0.004	0.002	0.000	0.416	0.242	0.128	0.210	0.141	0.078
S5	probit	−0.010	−0.090	−0.025	14.251	8.242	5.129	0.346	0.288	0.223
	KS1	2.716	3.137	3.541	3.503	3.973	4.391	2.594	2.861	3.147
	KS2	1.352	1.348	1.512	17.751	2.883	2.936	1.637	1.226	0.240
	SKS	−0.005	−0.001	0.001	1.015	0.393	0.145	0.233	0.143	0.078
S6	probit	0.012	0.022	0.010	1.282	2.430	0.797	0.346	0.282	0.225
	KS1	5.440	5.809	6.102	5.902	6.250	6.534	5.538	5.728	5.853
	KS2	3.180	3.067	3.260	7.778	6.077	5.860	5.303	4.796	1.939
	SKS	−0.118	−0.176	−0.047	4.767	4.222	2.114	2.830	1.084	0.090

Notes: "Start" refers to the specification of the starting values, and "KS1" and "KS2" refer to the KS estimator when the bandwidth is set equal to $h = n^{-1/6.02}$ and when it is estimated, respectively. The bias, RMSE and MAD are aggregated across regressor designs, error designs and success probabilities.

Table 2: Aggregated Monte Carlo results for E2–E5.

		Bias			RMSE			MAD		
Start	Estimator	$n = 250$	$n = 500$	$n = 1500$	$n = 250$	$n = 500$	$n = 1500$	$n = 250$	$n = 500$	$n = 1500$
	LPM	−0.104	−0.098	−0.094	5.358	0.820	0.659	0.388	0.315	0.255
S1	probit	−0.227	−0.259	0.038	74.461	72.089	103.070	0.491	0.397	0.292
	KS1	−0.016	−0.013	−0.008	0.368	0.247	0.134	0.129	0.085	0.042
	KS2	−0.019	−0.011	−0.004	0.963	0.469	0.252	0.266	0.199	0.122
	SKS	0.016	0.019	0.012	0.591	0.404	0.235	0.296	0.204	0.114
S2	probit	−0.095	−0.046	−0.190	47.089	38.663	297.030	0.441	0.349	0.252
	KS1	−0.093	−0.086	−0.081	5.349	0.768	0.600	0.359	0.264	0.165
	KS2	−0.067	−0.067	−0.062	7.791	0.908	0.717	0.338	0.241	0.143
	SKS	−0.009	0.001	−0.002	5.332	0.534	0.343	0.320	0.216	0.120
S3	KS1	−0.118	−0.065	−0.197	47.086	38.662	297.040	0.399	0.284	0.171
	KS2	0.129	−0.013	−0.182	61.797	49.442	297.320	0.370	0.257	0.146
	SKS	−0.065	−0.003	−0.165	47.108	38.804	297.020	0.335	0.224	0.123
S4	probit	0.185	0.424	1.347	59.389	526.180	1211.200	0.590	0.491	0.385
	KS1	0.799	0.840	0.868	1.641	1.602	1.665	0.921	0.805	0.596
	KS2	0.685	0.276	0.191	46.788	11.649	1.010	0.600	0.388	0.195
	SKS	0.023	0.025	0.021	0.806	0.494	0.274	0.339	0.223	0.124
S5	probit	−2.515	−0.477	−0.143	1188.600	775.840	79.839	0.577	0.487	0.388
	KS1	2.526	2.790	3.201	3.167	3.408	3.896	2.528	2.795	3.124
	KS2	2.754	1.479	1.516	81.124	28.706	4.541	1.950	1.646	0.845
	SKS	0.006	0.007	0.016	1.840	1.129	0.450	0.488	0.256	0.126
S6	probit	0.185	0.424	1.347	59.389	526.180	1211.200	0.590	0.491	0.385
	KS1	5.206	5.445	5.806	5.565	5.773	6.135	5.422	5.629	5.836
	KS2	4.582	3.289	3.448	76.951	8.626	6.627	5.341	5.246	4.007
	SKS	0.058	−0.129	−0.127	5.231	4.849	3.788	3.038	2.899	0.251

Notes: See Table 1 for an explanation.

Table 3: Empirical coefficient estimates based on KS.

	KS1				KS2			KS3			
Regressors	S3	S2	S5′	S6′	S3	S2	S6′	S3	S2	S5′	S6′
Risk aversion	-0.548^{**}	-0.533^{**}	-2.229^{***}	-0.980^{*}	-0.350^{***}	-0.348^{***}	-0.824^{*}	4.342^{***}	-0.873^{*}	-1.868^{***}	-2.200^{***}
	(0.261)	(0.260)	(0.287)	(0.522)	(0.048)	(0.044)	(0.463)	(0.767)	(0.455)	(0.480)	(0.560)
Ambiguity aversion	-0.194	-0.192	-0.355	0.111	-0.020	-0.022^{***}	1.064^{*}	-1.501^{***}	-0.255	0.526	0.533
	(0.278)	(0.278)	(0.335)	(0.440)	(0.036)	(0.036)	(0.554)	(0.320)	(0.380)	(0.415)	(0.430)
Age	0.158^{***}	0.160^{***}	0.380^{***}	0.598^{***}	0.094^{***}	0.097^{***}	0.700^{**}	-1.240^{***}	0.225^{***}	1.489^{***}	1.138^{***}
	(0.027)	(0.027)	(0.037)	(0.051)	(0.008)	(0.007)	(0.288)	(0.175)	(0.080)	(0.257)	(0.245)
Female	-0.022	-0.016	-0.113	0.418^{*}	0.119^{***}	0.120^{***}	1.415^{**}	-1.670^{***}	0.019	2.152^{***}	2.841^{***}
	(0.114)	(0.114)	(0.119)	(0.213)	(0.018)	(0.016)	(0.595)	(0.306)	(0.173)	(0.385)	(0.642)
German grade	0.102	0.102	0.289^{***}	-0.015	0.005	0.007	0.394^{**}	1.346^{***}	0.101	0.172^{*}	0.654^{***}
	(0.079)	(0.078)	(0.087)	(0.121)	(0.010)	(0.009)	(0.187)	(0.220)	(0.102)	(0.105)	(0.223)
Math grade	-0.321^{***}	-0.334^{***}	-0.814^{***}	0.880^{***}	-0.178^{***}	-0.181^{***}	-2.703^{**}	4.495^{***}	-0.434^{***}	2.494^{***}	-3.283^{***}
	(0.073)	(0.073)	(0.088)	(0.118)	(0.016)	(0.015)	(1.120)	(0.620)	(0.138)	(0.428)	(0.643)
No. of siblings	0.273^{***}	0.269^{***}	0.701^{***}	0.838^{***}	0.027^{***}	0.027^{***}	0.734^{**}	0.650^{***}	0.134	0.503^{***}	0.204
	(0.043)	(0.043)	(0.043)	(0.095)	(0.007)	(0.006)	(0.298)	(0.106)	(0.111)	(0.123)	(0.182)
Pocket money	0.001	0.001	-0.003	0.009^{**}	-0.002^{**}	-0.002^{**}	0.072^{**}	0.060^{***}	-0.005	-0.065^{***}	-0.033^{***}
	(0.002)	(0.002)	(0.002)	(0.004)	(0.001)	(0.001)	(0.031)	(0.009)	(0.005)	(0.011)	(0.010)

Notes: "KS1" and "KS2" refer to the KS estimator when the bandwidth is set equal to $h = n^{-1/6.02}$ and when it is estimated, respectively. KS1 and KS2 are implemented in Matlab. "KS3" refers to the Stata implementation of KS. "S2", "S3", "S5′" and "S6′" refer to different starting value specifications. KS2 failed to converge when initialized based on S5′. The corresponding column is therefore omitted. The coefficient of delay aversion is normalized to 1 in the estimation and is therefore not reported. The numbers within parentheses are the estimated standard errors. ***, ** and * signify significance at the 1%, 5% and 10% level, respectively.

Table 4: Empirical coefficient estimates based on LPM, probit and SKS.

	SKS					
Regressors	S3	S2	S5′	S6′	LPM	Probit
Risk aversion	0.038	0.038	0.038	0.038	−0.025	−0.014
	(0.063)	(0.063)	(0.063)	(0.063)	(0.262)	(0.275)
Ambiguity aversion	−0.050	−0.050	−0.050	−0.050	−0.066	−0.065
	(0.060)	(0.060)	(0.060)	(0.060)	(0.247)	(0.254)
Age	0.014**	0.014**	0.014**	0.014**	0.102**	0.089*
	(0.007)	(0.007)	(0.007)	(0.007)	(0.051)	(0.046)
Female	0.040	0.040	0.040	0.040	0.065	0.091
	(0.029)	(0.029)	(0.029)	(0.029)	(0.119)	(0.124)
German grade	−0.017	−0.017	−0.017	−0.017	−0.006	−0.016
	(0.018)	(0.018)	(0.018)	(0.018)	(0,075)	(0.078)
Math grade	−0.015	−0.015	−0.015	−0.015	−0.258*	−0.251*
	(0.017)	(0.017)	(0.017)	(0.017)	(0.133)	(0.134)
No. of siblings	0.017	0.017	0.017	0.017	0.111	0.070
	(0.016)	(0.016)	(0.016)	(0.016)	(0.082)	(0.070)
Pocket money	−0.001	−0.001	−0.001	−0.001	−0.002	−0.002
	(0.001)	(0.001)	(0.001)	(0.001)	(0.004)	(0.003)

Notes: Probit failed to converge when initialized based on S5′ and S6′. The results in the table are therefore based on LPM initialization (S2). See Table 3 for an explanation of the rest.

Table 5: Empirical marginal effect estimates based on KS and SKS.

Regressor	Mean	Min	Quartile 1	Median	Quartile 3	Max
			KS			
Risk aversion	0.033	−0.020	0.003	0.034	0.050	0.161
Ambiguity aversion	0.012	−0.007	0.001	0.012	0.018	0.057
Age	−0.010	−0.046	−0.014	−0.010	−0.001	0.006
Female	0.001	−0.001	0.000	0.001	0.002	0.007
German grade	−0.006	−0.030	−0.009	−0.006	−0.001	0.004
Math grade	0.019	−0.012	0.002	0.020	0.029	0.094
No. of siblings	−0.016	−0.080	−0.025	−0.017	−0.002	0.010
Pocket money	0.000	0.000	0.000	0.000	0.000	0.000
			SKS			
Risk aversion	0.032	0.023	0.031	0.033	0.034	0.034
Ambiguity aversion	−0.042	−0.045	−0.044	−0.043	−0.040	−0.030
Age	0.012	0.008	0.011	0.012	0.013	0.013
Female	0.033	0.024	0.032	0.034	0.035	0.036
German grade	−0.014	−0.015	−0.015	−0.014	−0.013	−0.010
Math grade	−0.012	−0.013	−0.013	−0.013	−0.012	−0.009
No. of siblings	0.014	0.010	0.013	0.014	0.015	0.015
Pocket money	−0.001	−0.001	−0.001	−0.001	−0.001	−0.001

Notes: The table reports summary statistics of the distribution of the estimated marginal effects for every regressor over the entire sample.