Oberrauch, Luis; Kaiser, Tim; Seeber, Günther

Working Paper

Measuring Economic Competence of Youth with a Short Scale

Measuring Economic Competence of Youth with a Short Scale

Luis Oberrauch, Tim Kaiser, and Günther Seeber

Abstract

We present a 12-item scale measuring the cognitive component of economic competence and document the psychometric properties of the scale. Using a data set with more than 12,000 secondary school students in Germany, the scale shows high discriminatory power and covers a wide range of ability levels. Analyses of `Differential Item Functioning´ show no item bias across key demographic characteristics, and scores show meaningful associations with scores obtained from adjacent test instruments. Student-level correlates mirror estimates documented in earlier literature on economic and financial literacy as well as results relying on a more extensive scale with over 30 items. The presented short scale enables researchers and practitioners to efficiently measure economic competence of youth.

JEL-Classification: A21, G53, I21
Keywords: economic competence, economic literacy, IRT, measurement

March 2022

Acknowledgments: We thank the schools and students participating in the assessments. Research assistance by Ngoc Anh Nguyen and financial support by *Stiftung Würth* are greatly appreciated. Supplemental replication data and code can be accessed via the Open Science Framwork under https://osf.io/cn8t7/ or DOI 10.17605/OSF.IO/CN8T7.

Luis Oberrauch, Eberhard Karls University Tuebingen, D-72074 Tübingen, Germany; luis.oberrauch@uni-tuebingen.de

Tim Kaiser (corresponding author), University of Koblenz-Landau, D-76829 Landau, Germany (ORCiD: https://orcid.org/0000-0001-7942-6693); kaiser@uni-landau.de

Günther Seeber, University of Koblenz-Landau, D-76829 Landau; seeber@uni-landau.de

1 Introduction

Research on the economic understanding and behavior of children and youth has a long and ongoing tradition in economics and psychology (e.g., Strauss 1952, Danziger 1958, Berti and Bombi 1981, Leiser 1983, Furnham and Bond 1986, Furnham and Cleare 1988, Sevon and Weckstrom 1989, Leiser and Halachmi 2006, Davies and Lundholm 2012, Grohmann et al. 2015, Lührmann et al. 2015, Sutter et al. 2019, Brocas et al. 2019, Andreoni et al. 2020, Brocas and Carillo 2020, Choshen-Hillel et al. 2020, Brocas and Carillo 2021). An integral part of many of these empirical inquiries is the measurement of knowledge and skills in the economic domain. In recent years, research on financial literacy, i.e., a subset of general economic literacy focused on individual financial decision-making competencies, has gained increased interest with many empirical studies relying on few test items to derive financial literacy scores (see Hastings et al. 2013, Lusardi and Mitchell 2014, Kaiser et al. 2021 for reviews of the literature). In contrast, research on economic literacy and related constructs (i.e., knowledge and skills in the broader economic domain) has relied on more extensive measurement scales targeted at different audiences: The most established measurement scale targeted at U.S. high school students, the *Test of Economic Literacy* (e.g., Walstad et al. 2013) is comprised of 45 items, and the most widely known test measuring economic understanding of college students (Walstad and Rebeck 2008) encompasses 60 items. While the use of these elaborate measurement instruments potentially allows a precise measurement of the underlying latent constructs, they also come at the cost of a substantial response burden for students rendering the implementation in surveys not primarily geared towards a single objective unlikely.

To address this gap, we present a short (i.e., 12-item) scale to efficiently measure the cognitive component of *economic competence*, i.e., problem solving capability in the economic domain. We select items from the long form *Test of Economic Competence* (TEC) (Kaiser et al. 2020) based on psychometric properties and applicability for English-speaking respondents.

Using a large sample of 12,146 school students from Germany, we analyze construct validity (and equivalence) and item characteristics of the short scale using Item Response Theory (IRT). Additionally, we investigate associations to adjacent constructs relevant to economic decision-making as well as student and group-level correlates relative to results from prior literature.

We present four findings: First, fit statistics show the 12-item scale is unidimensional, i.e., measuring a single latent construct. This is particularly relevant in the economic domain, where contents and cognitive processes may overlap with other domains, such as civic education or mathematics. Second, regarding characteristics of single items, estimates based on Classical Test Theory (CTT) and Item response theory (IRT) show that the items capture the underlying construct across a wide range of ability levels and are good discriminators between high and low-ability respondents. Third, we detect no item bias across three demographic characteristics (gender, native language, and socio-economic status) providing further evidence on the scale's construct validity and overall test fairness. Fourth, student- and group-level correlates mirror results documented in prior literature on economic knowledge and skills (e.g., Walstad et al. 2013, Oberrauch und Kaiser 2020) as well as financial literacy (e.g., Grohmann et al. 2015; Driva et al. 2016, Lührmann et al. 2015). Further evidence on criterion validity of the scale stem from correlations with adjacent constructs relevant for economic and financial decision-making. Specifically, test scores are positively associated with interest in economic matters, financial planning, and the propensity to save. By contrast, test scores are negatively correlated with impulse purchasing.

The paper addresses the lack of a widely disseminated short measure of economic competence by providing evidence on the construct validity of a short scale as well as evidence of the construct equivalence to its long-form version (Kaiser et al. 2020). We contribute to the literature by providing a test instrument enabling researchers to efficiently measure economic

competence in educational large-scale assessments and evaluations of treatment effects of educational interventions.

2 Conceptual model and content validity

The 12-item scale is selected from the long-form *Test of Economic Competence* containing 31 items (Kaiser et al. 2020). The scale is based on a conceptual model of economic competence, which defines economic competence as the ability of individuals to solve problems in three life situations, i.e., as (i) consumers of goods and services, (ii) employees and self-employed, (iii) as well as citizens. In each of these situations, students need to apply competences in three domain-specific areas referring to overall goals of general education (Kaiser et al. 2020, 230). Specifically, students A) make rational choices by taking constraints into account (*decision-making and rationality*), B) recognize and consider (economic) interests of other individuals, and C) understand economic mechanisms at the systemic and societal level. Thus, the model defines economic competence as a broad concept not exclusively focusing on individual decision-making but also on understanding institutions and systemic features of the economic system which is in line with previous work on normative competence goals discussed in the existing literature (e.g., Davies 2015).

Note that the three competence areas are content-based theoretical delimitations of a global competence construct instead of dimensions of their own, with A) representing the individual, B) interpersonal, and C) the systemic/societal perspective. Combining competence areas with the situations leads to a matrix defining competences at different levels of aggregation (e.g., Retzmann and Seeber 2016). Accordingly, items ought to measure whether students are able to analyze and evaluate the consequences of an economic decision (individual), cooperation (social), and institutions and policies from an economic perspective (systemic/societal) in context of the life situations. Aside from relying on the conceptual model

in the design of the items, further evidence on the content validity came from expert validations and think-aloud studies on each item (see Kaiser et al. 2020).

3 Data

The data consist of cross-sections of students in class-levels 7 to 10 in schools in the German federal state Baden-Wuerttemberg. The test and the survey capturing demographic, behavioral and attitudinal characteristics were administered during regular lessons and supervised by the respective teachers (Kaiser et al. 2020). The sampling process for all cross-sections followed the same procedure: We divided the population of students in schools in the relevant grade into strata based on school type and degree of urbanization). Within the strata (four school types across three degrees of urbanization) we followed a two-stage procedure. First, we randomly selected schools that would yield a similar proportion of students in the stratum as in the whole population of interest. Second, we randomly choose one class per school in the relevant class-level. To account for overrepresentation of larger (or smaller) schools, we use school size as an implicit stratification variable: We size schools within each stratum and deploy systematic sampling using sampling intervals. Remaining disproportionalities are addressed by the use of sampling weights (i.e., the inverse of the selection probability).

The sample consists of 12,146 students from 616 classes in 305 schools. Summary statistics are shown in Table 1 but for brevity are not discussed here. In the 10^{th} grade sample, we additionally captured behavioral and attitudinal variables (for summary statistics, see Table A1 in Appendix A).

Table 1: Sample characteristics

Variable	N	Mean	SD	Median	Min	Max
Individual characteristics						
Male	12,146	0.520			0	1
Age (years)	12,020	15.116	1.284	15	11	23

Other language at home	11,520	0.370			0	1
≤ 25 books at home	11,484	0.284			0	1
Reading abilities	12,136	3.868	0.748	4	1	5
Math abilities	12,124	3.437	0.969	3	1	5
Own bank account	11,376	0.715			0	1
Own ATM card	11,379	0.540			0	1
Own salary	11,436	0.725			0	1
Effort score	12,146	94.612	12.295	100	0	100
Class-level						
7th grade	12,146	0.251			0	1
8th grade	12,146	0.246			0	1
9th grade	12,146	0.315			0	1
10th grade	12,146	0.188			0	1
School-level characteristics						
Higher track school	12,146	0.371			0	1
Low urbanization	12,134	0.333			0	1
School size	12,146	604.92	217.53	606	111	1328

Notes: *Male* indictates the gender of repsondents. Age is student age in years. *Other language at home* is a dummy variable indicating whether students primarily speak a language other than german at home. We proxy students' socio-economic status by asking how many books there are in their home (excluding text books and magazines) on a scale from 1 (none) to 6 (several bookshelves) and create a dummy variable indicating ≤ *25 books at home*. *Reading and math-abilities* are self-reported by students on a scale from 1 (very low) to 5 (very high). We measure students' effort using the time-response effort approach (Wise and Kong 2005). Accordingly, a student exhibits effort on a particular item if the time spent exceeds the normative threshold of 10 percent of the mean response time. The final *effort score* represents the percentage of exhibited effort across all competence items. *Class-level* indicates the class-level of sampled students. *Higher-track school* indicates that the student visits the most sophisticated school type ("Gymnasium"). *Low urbanization* reports whether the student comes from a rural area. School size reports the number of students in the respective school.

4 Methods

Item Response Theory (IRT) models have been widely used in large-scale assessments (e.g., PISA or TIMMS) and aim to overcome shortcomings of Classical Test Theory (CTT) (see Baker and Kim 2004). While the CTT approach relies solely on observed scores (i.e., the number of solved items), IRT models assume the probability of endorsing an item to be a monotonically increasing function of an underlying trait (denoted by θ). In Figure A1 in Appendix A, this relationship is represented by item characteristic curves with their archetypal s-shape. The logistic function may depend on the item difficulty (i.e., the location of the curve), discrimination (i.e., the slope of the curve), guessing among low-ability examinees (i.e., the lower asymptote of the curve), and exhaustion or inattention among high-achieving examinees (i.e., the upper asymptote of the curve). Thus, IRT models allow investigating several item characteristics in the context of performance tests.

One assumption for the use of IRT models is that the underlying trait is unidimensional, i.e., the defined latent trait is the only construct to be inferred from observed item responses. Another requirement for the use of IRT is the local independence between the items, i.e., item responses must only correlate due to the underlying ability.

Further, educational large-scale assessments typically employ a variety of IRT models, ranging from one to four parameters: The general form of the unidimensional IRT model with four parameters (Magis 2013) is expressed as

$$P\left(\mathrm{X}_j = 1 \mid \theta_v, \sigma_i, \alpha_i, \gamma_i, \delta_i\right) = \gamma_i + (\delta_i - \gamma_i)\, \frac{\exp\left[\alpha_i(\theta_v - \sigma_i)\right]}{1 + \exp\left[\alpha_i(\theta_v - \sigma_i)\right]} \quad (1),$$

where θ_v denotes the ability of person v and σ_i the difficulty of item i on a logit scale. α_i represents the discrimination parameter, i.e., it measures how well the item distinguishes between low and high-ability individuals. γ_i is a guessing parameter measuring the probability of endorsing an item by low ability individuals (i.e., the lower asymptote of the Item Characterstic Curve). γ_i is the upper asymptote modelling the possibility of lower probabilities of endorsing an item due to fatigue or inattention. Prior to estimating model parameters, we assess which (unidimensional) IRT model for binary items (ranging from one to four parameters) best fits the data.

5 Results: Construct validity, convergent validity, and criterion validity

5.1 Model selection and modelling assumptions

Dimensionality. A common way to empirically assess the dimensionality of an item set is factor analysis. A principal component analysis on the item set revealed an eigenvalue of 3.93 (i.e., explaining about 16 % of the total variance) for the first factor while the eigenvalue for the second factor was almost four times smaller, with all remaining factors below 1. Consequently, these results indicate the dominance of the first factor and therefore provide evidence on the unidimensionality of the scale.

Model selection. Comparing chi-square fit statistics ($S-\chi^2$) (Orlando and Thissen 2003) across four model specifications (1-PL to 4-PL IRT models) reveals eight significant deviations for the 1-PL model, three deviations for the 2-PL model, one deviation for the 3-pl-model and no deviations for the four-parameter model ($p<0.01$) (for detailed results, see Table A1 in Appendix A). Thus, the IRT model described in equation (1) with four parameters appears to be the best fit to the 12-item short scale.

Local independence. Local independence implies that item responses must only correlate due to the underlying ability. Thus, we empirically probe the local independence assumption by keeping ability levels constant and examining Q3 statistics (Yen 1984). This revealed a mean residual correlation between item pairs of -0.063 (SD= 0.02), i.e., providing strong evidence for the local independence assumption to be fulfilled.

5.2 Item analysis

Table 2 shows psychometric properties of the scale based on CTT. Within this framework, item-total correlations (ITC) are one of the most important statistics and refers to the discriminatory power of an item. They represent point-biserial correlations between endorsing an item and the test score based on the remaining items. Positive coefficients indicate that high-achieving students are more likely to endorse the item while coefficients close to 0 or below 0 indicate that the item discriminates poorly between low and high-achieving students. Table 2 shows that all items are moderately but positively correlated with total test scores, indicating that the scale is functioning properly regarding discriminatory power. Frequency reports the percentage of correctly solved items and refers to item easiness in a CTT context. The results show that the item set covers a broad range of difficulties, with item 2 being the easiest and item 12 being the hardest item.

Table 2: Percentages of correct responses and item discrimination

Itemno.	r_{it}	Frequency (all)	Frequency (with mandate)	Frequency (without mandate)	Competence area
1	0.340	0.731	0.805	0.674	A
2	0.206	0.835	0.847	0.825	A
3	0.254	0.749	0.787	0.720	B
4	0.402	0.475	0.509	0.449	A
5	0.326	0.701	0.728	0.681	A
6	0.310	0.776	0.794	0.763	C
7	0.285	0.580	0.589	0.573	C
8	0.323	0.659	0.679	0.643	A
9	0.287	0.561	0.604	0.528	C
10	0.320	0.456	0.480	0.438	B
11	0.289	0.320	0.338	0.306	A
12	0.213	0.294	0.310	0.282	C

Notes: This table shows (corrected) item-total correlations r_{it} as well as percentages of correctly solved items (Frequency) for the whole sample. Columns 3 and 4 show frequencies for a subsample of 9th graders (*N=3,047*), with one group being exposed to mandatory economic education and one group not being exposed to mandatory economic education. The last column displays item contents regarding the competence areas delineated in the conceptual competence model: "A: decision-making and rationality", "B: relationships and interaction" and "C: system and order".

Additional evidence on the construct validity of the scale is presented in Columns 4 and 5. We expect that students who have had mandatory economic education in school exhibit higher test scores relative to students without mandatory economic education. To test this hypothesis, we rely on a subsample of 9th graders: Half of this subsample was surveyed in 2019 prior to a policy reform introducing mandatory economic education (see Kaiser and Oberrauch 2021) while the other half was surveyed in 2020 and received two to three years of instruction depending on school type. Our results show the percentages of endorsed items are higher across the entire item for students covered by the mandate, i.e., further providing evidence that the latent construct is captured by the scale. Next, estimated parameters of the IRT model specified in equation (1) are shown in Table 4. In essence, results from the four-parameter IRT model mirror CTT results: We observe a broad variation in the difficulty parameter $\hat{\sigma}$ ranging from -1.733 (Item 2) to 1.597 (Item 12) on the logit scale. Following guidelines for the interpretation of the parameter $\hat{\alpha}$ (e.g., Baker and Kim 2004), most items show high to very high discriminatory power ($\hat{\alpha}$ >1.35), while items 2, 3 and 9 appear to be moderate discriminators.

Further, we estimate item information curves describing the amount of information an item provides at various points of the trait continuum (see Figure A1). Additionally, Figure A2 in Appendix A shows the test information curve, i.e., the sum of the individual item information curves, of the 12-item scale. The test information curve reaches its maximum at $\theta = 0.16$ and is well distributed indicating that the scale can reliably measure a broad range of competence levels.

Table 3: Four-parameter IRT estimates and model fit statistics

Itemno.	$\hat{\alpha}$ [SE]	$\hat{\sigma}$ [SE]	$\hat{\gamma}$ [SE]	$\hat{\delta}$ [SE]	$S-\chi^2$ (p-val.)
1	3.388 [0.189]	-0.110 [0.022]	0.451 [0.012]	0.982 [0.003]	6.308 (0.504)
2	0.968 [0.022]	-1.733 [0.032]	0.000 [0.019]	0.936 [0.005]	16.124 (0.024)
3	0.902 [0.019]	-1.068 [0.025]	0.000 [0.013]	0.996 [0.005]	12.879 (0.075)
4	1.983 [0.068]	0.272 [0.020]	0.129 [0.009]	1.000 [0.005]	5.573 (0.590)
5	3.435 [0.148]	0.062 [0.015]	0.303 [0.007]	0.934 [0.005]	6.838 (0.446)
6	1.405 [0.032]	-1.004 [0.020]	0.000 [0.012]	0.973 [0.004]	4.278 (0.747)
7	1.367 [0.062]	-0.171 [0.032]	0.142 [0.013]	0.898 [0.009]	6.521 (0.480)
8	1.478 [0.043]	-0.002 [0.020]	0.177 [0.008]	1.000 [0.006]	4.267 (0.749)
9	1.097 [0.049]	0.285 [0.034]	0.156 [0.011]	1.000 [0.011]	10.649 (0.155)
10	1.878 [0.055]	0.589 [0.018]	0.078 [0.005]	0.892 [0.010]	5.994 (0.540)
11	2.164 [0.109]	1.200 [0.027]	0.179 [0.008]	1.000 [0.013]	7.024 (0.426)
12	3.357 [0.236]	1.507 [0.027]	0.240 [0.007]	1.000 [0.009]	15.186 (0.034)

Notes: This table shows estimated parameters and standard errors (in brackets) based on the four-parameter IRT model displayed in equation (1). $\hat{\alpha}$ denotes the discrimination parameter, $\hat{\sigma}$ represents item difficulty (i.e., location of the ICC), $\hat{\gamma}$ denotes the guessing parameter, whereas $\hat{\delta}$ reflects inability among high-ability respondents. Column 6 reports fit indices based on the $S-\chi^2$approach with corresponding p-values (in parenthesis).

5.2 Differential Item Functioning

As described in section 4, the IRT model assumes measurement invariance, i.e., estimated parameters are the same regardless of which demographic subgroup (e.g., gender or mother tongue) respondents belong to. Measurement variance indicates that an additional construct is measured by the item potentially violating the unidimensionality assumption. To detect item-level bias, analysis of (uniform) Differential Item Functioning (DIF) is commonly employed in educational assessments (Holland and Wainer 2012; Thissen et al. 1993): By following the MH method (Mantel and Haenszel 1959), we compute the magnitude of DIF α_{MH} by scaling the subgroups separately and then test them for statistical differences. The significance is tested using the Mantel-Haenszel chi-square test. To ease the interpretation of

the MH statistic α_{MH} , Holland and Thayer (1986) proposed a transformed index that can be interpreted as effect size and is simply defined as Δ_{MH} = -2.35 α_{MH} (Δ −index). Depending on the severity of DIF, this approach classifies DIF into three categories: $|\Delta_{MH}| \leq 1$ denotes no or negligible DIF (category A), $1 < |\Delta_{MH}| \leq 1.5$ corresponds to moderate DIF (category B) and $|\Delta_{MH}| > 1.5$ denotes severe DIF (category C). We use three demographic split criteria that have shown to be predictive for test scores in previous studies (e.g., Oberrauch and Kaiser 2020): Gender, native language, and the number of books at home. As shown in Table A3 in the Appendix, only two items show moderate DIF regarding gender.

Item no. 2 moderately disadvantages the focal group (female) while item no. 10 moderately disadvantages the reference group, i.e., male respondents. For the demographic criterions socio-economic status, proxied by the number of books at home, and native language, all items show merely negligible DIF. Overall, the results provide further evidence on the construct validity and the test fairness with respect to three key demographic characteristics.

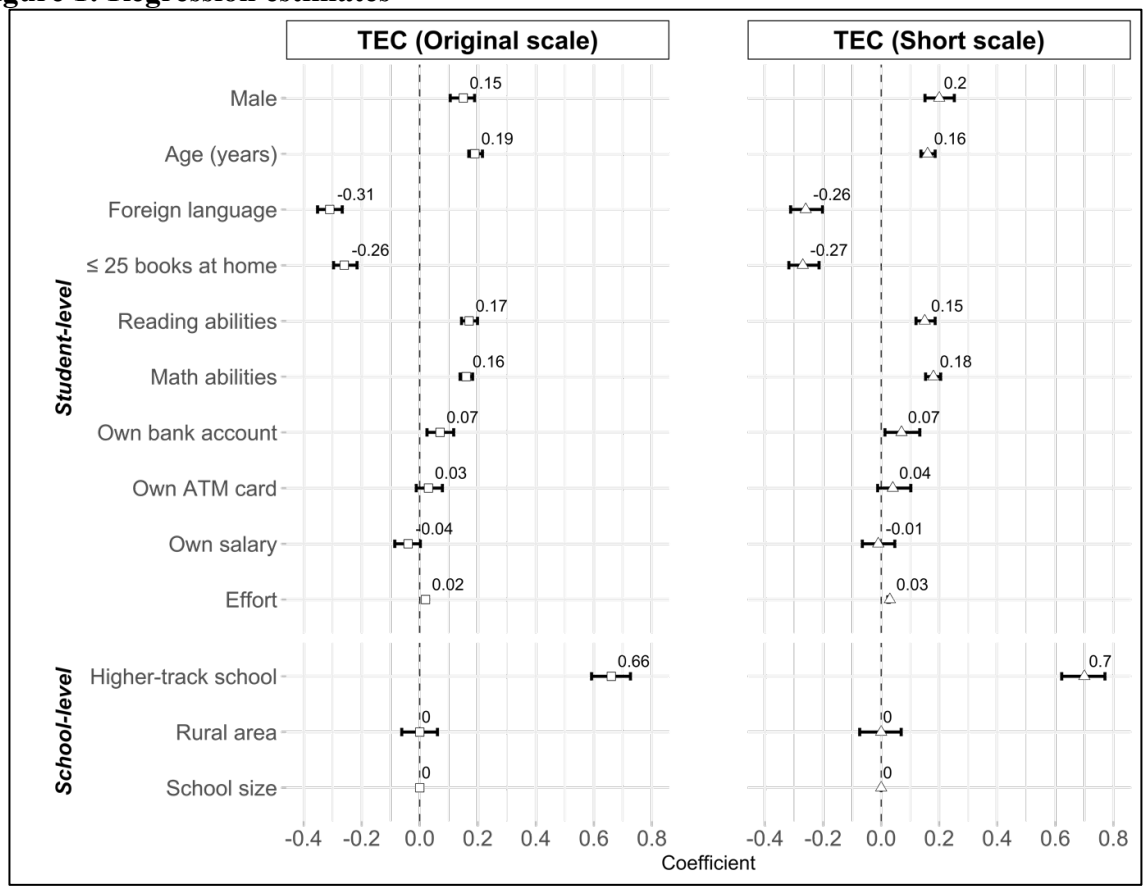
5.3 Convergent validity

To assess the convergent validity of the item set, we analyze correlations to adjacent capability measures in the economic domain. In addition to available test scores relying on the long form scale with 31 items (*N=12,146*), we administered a common economic knowledge test (Eberle and Oberrauch 2022) to a subsample of 9th graders (*N=2,843*). The knowledge test relies on content-oriented knowledge including questions about unemployment in Germany and about the legal age to take up a loan. The results show a positive correlation with scores on the long form test ($r=0.845$, $p<0.01$) and with knowledge scores ($r=0.48$, $p<0.01$) (see Figure A3 in Appendix A).

5.4 Criterion validity

Next, we investigate correlations between socio-demographic characteristics with test scores and compare them to correlations with test scores obtained from long from test (see Figure 1). Point estimates on the 12-item set scores mirror estimates relying on the original scale with 31 items. We observe a positive association between scores and being male, age (in years), reading and math abilities, test effort as well as enrollment in higher track schools. The results correspond with findings from previous literature (e.g., Grohmann et al. 2015; Lührmann et al. 2015; Oberrauch and Kaiser 2020; Kaiser et al. 2020). However, slightly wider confidence intervals using the short scale indicate less precision in estimates compared to results obtained from the original scale.

Figure 1: Regression estimates



Notes: OLS regression coefficients are displayed with 95% CIs. Dependent variables are IRT scores obtained from the original scale with 31 items (complete set) and from the 12-item short scale (reduced set). To account for measurement error, we used 20 plausible values based on a basic latent regression model. Dependent variables are z-standardized to have a mean of 0 and a standard deviation of 1. Non-categorial variables are mean-centered. Number of observations are n=9,453 (complete set) and n=9,452 (reduced set). Adjusted R^2 are 0.4 and 0.32, respectively. Standard errors are clustered at the class-level.

Next, we investigate correlations with attitudinal and behavioral outcomes related to economic decision-making using a subsample of 10th-graders: Economic interest, financial planning, attitudes towards money, financial autonomy, impulse purchasing, and whether the respondent has any savings. The first four outcomes are multi-item scales whereas the last two outcomes (impulse purchasing and savings) are measured via single items. All scales and items are described in Appendix C.

Table 4 reports correlations of competence scores obtained from the short scale (Panel A) and from the long form scale (Panel B) with the above-mentioned outcomes. Again, correlations with scores obtained from the short 12-item scale (Panel A) mirror results relying on the original scale (Panel B). Competence scores are positively associated with economic interest, financial planning, financial autonomy, and the propensity to save. Expectedly, scores are negatively associated with the propensity to purchase impulsively whereas no correlation with attitudes towards money exists.

Table 4: Bivariate correlations with external constructs

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Correlations with competence scores relying on the 12-item short scale (*N=1,287*)							
(1) Economic competence	---						
(2) Economic interest	0.15***	---					
(3) Financial planning	0.07**	0.27***	---				
(4) Attitude towards money	-0.01	0.19***	0.12***	---			
(5) Financial autonomy	0.16***	0.20***	0.23***	0.08**	---		
(6) Impulse purchases	-0.17***	-0.10***	-0.42***	0.09**	-0.21***	---	
(7) Any savings	0.25***	0.09**	0.15***	0.00	0.17***	-0.08**	---
Panel B: Correlations with competence scores relying on the 31-item original scale (*N=1,286*)							
(1) Economic competence	---						
(2) Economic interest	0.19***	---					
(3) Financial planning	0.09***	0.27***	---				
(4) Attitude towards money	-0.02	0.19***	0.12***	---			
(5) Financial autonomy	0.19***	0.20***	0.23***	0.08**	---		
(6) Impulse purchases	-0.20***	-0.10***	-0.42***	0.09**	-0.21***	---	
(7) Any savings	0.26***	0.09**	0.15***	0.00	0.17***	-0.08**	---

Note: This table reports bivariate correlations based on Pearson's correlations as well as point-biserial correlations between competence, attitude and behavior scales using pairwise complete observations. * p<.05 ** p<.01 *** p<.001.

5.5 Sensitivity analyses

We conducted several sensitivity analyses, with methods and results described in Appendix B. First, we probe for Nonuniform Differential Item Functioning (NDIF) (Swaminathan and Rogers 1990) assuming the difference (relative to the focal group) in the probability of endorsing an item varies depending on ability levels within the subgroup, i.e., resulting in non-parallel item characteristic curves. The results show that all items of the short scale are flagged for negligible (nonuniform) DIF. Second, to support evidence of a unidimensional scale, we estimate item fits for alternative multidimensional models. Likelihood ratio tests show that the unidimensional model (Eq. 1) fits the data better than the considered multidimensional models.

6 Conclusion

This paper presented a short 12-item scale for measuring economic competence, i.e., problem solving capability in the economic domain. The items address a wide range of ability levels and appear to be good discriminators between high-achieving and low-achieving students. The analysis revealed no meaningful DIF effects across key demographic characteristics ensuring that demographic correlates with test scores are independent from potential item bias. Estimated differences in test scores across individual and school-level characteristics correspond with results already documented in the adjacent literature and with results relying on the original long-form scale. Further, test scores appear to be significantly correlated with constructs relevant for economic decision-making, such as financial planning, financial autonomy, and interest in economic matters.

Collectively, the results provide evidence on the construct and criterion validity of the short scale. As educational large-scale assessments aspire to capture competences rather than knowledge, we hope to provide researchers an efficient tool to be implemented in educational surveys and impact evaluations.

References

Andreoni, J., Di Girolamo, A., List, J. A., Mackevicius, C., & Samek, A. (2020). Risk preferences of children and adolescents in relation to gender, cognitive skills, soft skills, and executive functions. *Journal of Economic Behavior & Organization*, *179*, 729-742.

Baker, F. B., & Kim, S.-H. (2004). *The Basics of Item Response Theory Using* R. Springer.

Berti, A. E., & Bombi, A. S. (1981). The development of the concept of money and its value: A longitudinal study. *Child Development*, 52, 1179–1182.

Brocas, I., Carrillo, J. D., Combs, T. D., & Kodaverdian, N. (2019). The development of consistent decision-making across economic domains. *Games and Economic Behavior*, *116*, 217-240.

Brocas, I., & Carrillo, J. D. (2020). The evolution of choice and learning in the two-person beauty contest game from kindergarten to adulthood. *Games and Economic Behavior*, *120*, 132-143.

Brocas, I., & Carrillo, J. D. (2021). Steps of reasoning in children and adolescents. *Journal of Political Economy*, *129*(7), 2067-2111.

Choshen-Hillel, S., Lin, Z., & Shaw, A. (2020). Children weigh equity and efficiency in making allocation decisions: Evidence from the US, Israel, and China. *Journal of Economic Behavior & Organization*, *179*, 702-714.

Danziger, K. (1958). Childrens earliest conceptions of economic relations. *Journal of Social Psychology*, 47, 231–240.

Davies, P., & Lundholm, C. (2012). Students' understanding of socio-economic phenomena: Conceptions about the free provision of goods and services. *Journal of Economic Psychology*, *33*(1), 79-89.

Davies, P. (2015). Towards a framework for financial literacy in the context of democracy. *Journal of Curriculum Studies*, *47*(2), 300-316.

Driva, A., Lührmann, M., & Winter, J. (2016). Gender differences and stereotypes in financial literacy: Off to an early start. *Economics Letters*, 146, 143-146.

Eberle, M., & Oberrauch, L. (2022). What a difference three years of economics education make: Evidence from lower-track schools in Germany. ZBW Leibniz Information Centre for Economics.

Furnham, A., & Bond, M. (1986). Hong Kong Chinese explanations for wealth. *Journal of Economic Psychology*, 7(4), 447-460.

Furnham, A., & Cleare, A. (1988). School childrens conceptions of economics: Prices, wages, investment and strikes. *Journal of Economic Psychology*, 9, 467–479.

Grohmann, A., Kouwenberg, R., & Menkhoff, L. (2015). Childhood roots of financial literacy. *Journal of Economic Psychology*, *51*, 114-133.

Hastings, J. S., Madrian, B. C., & Skimmyhorn, W. L. (2013). Financial literacy, financial education, and economic outcomes. *Annu. Rev. Econ.*, *5*(1), 347-373.

Holland, P. W., & Thayer, D. T. (1986). Differential Item Functioning And The Mantel-Haenszel Procedure. ETS Research Report Series 1986 (2), i-24.

Holland, P. W., & Wainer H. (2012). *Differential item functioning* . Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Kaiser, T., Oberrauch, L., & Seeber, G. (2020). Measuring economic competence of secondary school students in Germany. *Journal of Economic Education*, *51*(3-4), 227-242.

Kaiser, T., Lusardi, A., Menkhoff, L., & Urban, C. (2021). Financial education affects financial knowledge and downstream behaviors. *Journal of Financial Economics*, forthcoming.

Kaiser, T., & Oberrauch, L. (2021). Economic education at the expense of indoctrination? Evidence from Germany, ZBW - Leibniz Information Centre for Economics, Kiel, Hamburg.

Leiser, D., (1983). Children's conceptions of economics - The constitution of a cognitive domain. *Journal of Economic Psychology*, 4, 297-317.

Leiser, D., & Halachmi, R. B. (2006). Children's understanding of market forces. *Journal of Economic Psychology*, 27, 6–19.

Lührmann, M., Serra-Garcia, M., & Winter, J. (2018). The impact of financial education on adolescents' intertemporal choices. *American Economic Journal: Economic Policy*, *10*(3), 309-32.

Lusardi, A., & Mitchell, O. S. (2014). The economic importance of financial literacy: Theory and evidence. *Journal of Economic Literature*, *52*(1), 5-44.

Magis, D. (2013). A Note on the Item Information Function of the Four-Parameter Logistic Model. *Applied Psychological Measurement*, 37 (4), 304–315.

Mantel, N. & Haenszel, W. (1959). Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease. *Journal of the National Cancer Institute,* 22, 719–748.

Oberrauch, L., & Kaiser, T. (2020). Economic competence in early secondary school: Evidence from a large-scale assessment in Germany. *International Review of Economics Education*, *35*, 100172.

Orlando, M. & Thissen, D. (2003). Further investigation of the performance of S_X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement* 27 (4), 289–98.

Retzmann, T., and G. Seeber. (2016). Financial education in general education schools: A competence model. In *International handbook of financial literacy*, ed. C. Aprea, E. Wuttke, K. Breuer, N. K. Koh, P. Davies, and B. Greimel-Fuhrmann, 9–23. Singapore: Springer.

Sevon, G., & Weckstrom, S. (1989). The development of reasoning about economic events: A study of Finnish children. *Journal of Economic Psychology*, 10, 495–514.

Strauss, A. L. (1952). The development and transformation of monetary meanings in the child. *American Sociological Review, 17,* 275–286.

Sutter, M., Zoller, C., & Glätzle-Rützler, D. (2019). Economic behavior of children and adolescents–A first survey of experimental economics results. *European Economic Review*, *111*, 98-121.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement,* 27 (4), 361–370.

Thissen, D., L. Steinberg, & Wainer H. (1993). *Detection of differential item functioning using the parameters ofitem response models. In Differential item functioning*, ed. P. W. Holland and H. Wainer, 67–113. Hillsdale, NJ:Lawrence Erlbaum Associates.

Walstad, W. B., & Rebeck, K. (2008). The test of understanding of college economics. *American Economic Review*, 98(2), 547-51.

Walstad, W. B., Rebeck, K., & Butters, R. B. (2013). The test of economic literacy: Development and results. *Journal of Economic Education*, *44*(3), 298-309.

Yen, W. M. (1984): Effects of Local Item Dependence on the Fit and Equating Performance of the Three-Parameter Logistic Model. *Applied Psychological Measurement,* 8 (2), 125–145.

Supplementary material

(Online supplement not intended for print publication)

to accompany

"Measuring Economic Competence of Youth with a Short Scale"

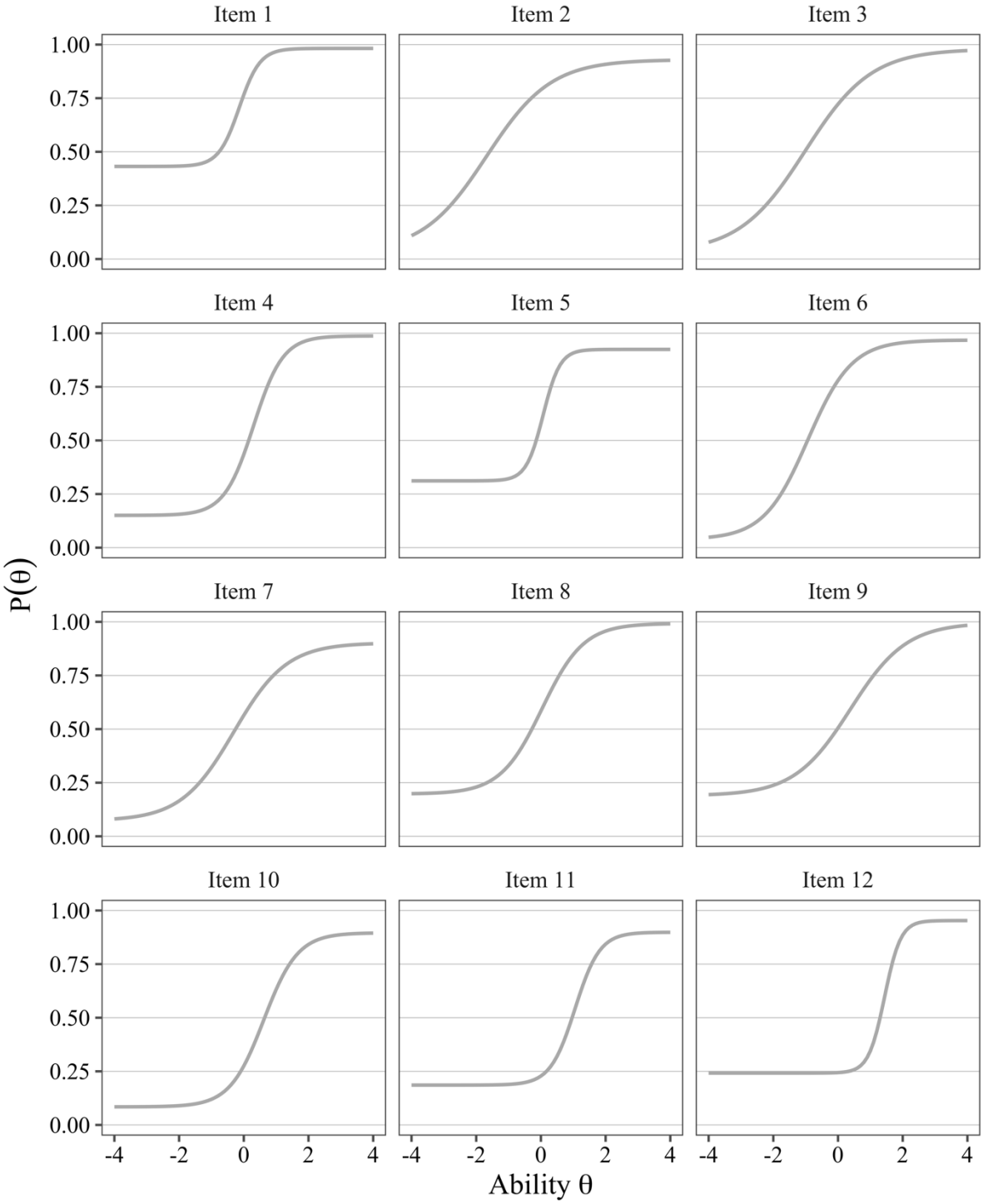
Appendix A: Auxiliary figures and tables

Appendix B: Sensitivity analyses

Appendix C: Items

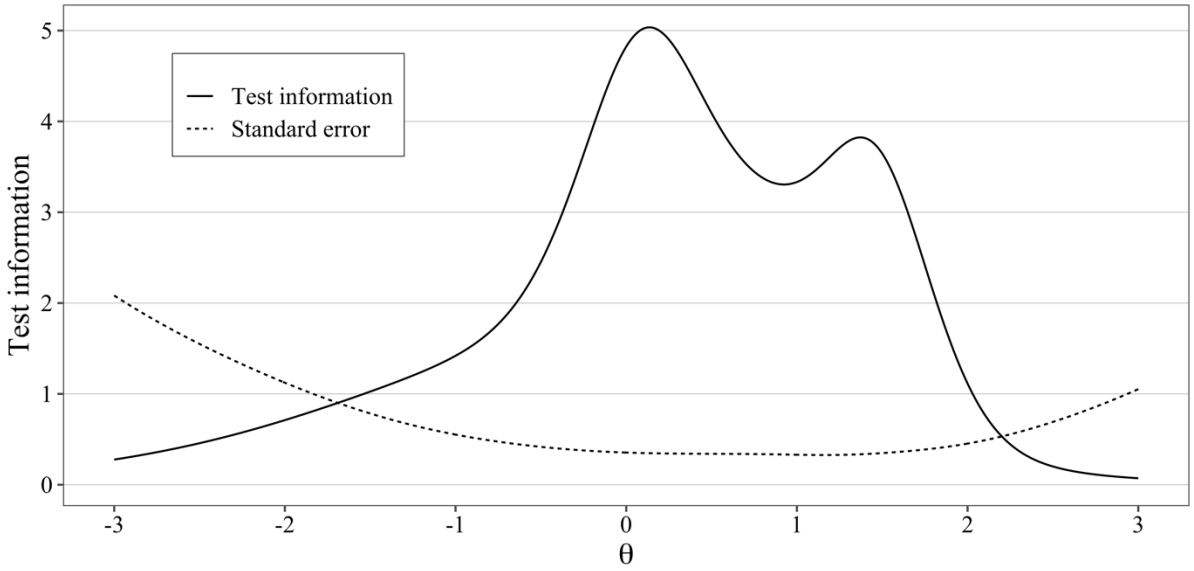
Appendix A: Auxiliary figures and tables

Figure A1: Item characteristic curves



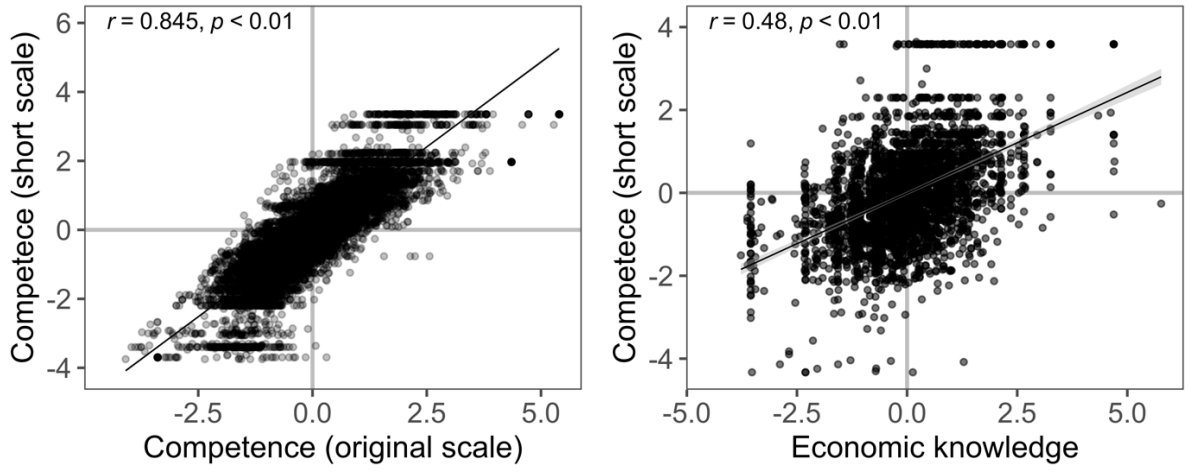
Notes: This figure shows Item Characteristic Curves for all 12 items of the short scale based on the four-parameter model displayed in equation (1) and described in chapter 4.

Figure A2: Test information curve



Notes: The solid line represents the test information as a function of the latent trait. The dotted line represents its standard error, calculated by the inverse of the item information's square root.

Figure A3: Correlations of test scores with scores on adjacent scales



Notes: This figure reports bivariate correlations (Pearson's r) between scores obtained from the short scale with scores obtained from the original scale (left panel), as well as with scores obtained from the economic knowledge scale described in chapter 5.3.

Table A1: Auxiliary summary statistics

Variable	N	Mean	SD	Median	Min	Max
Attitudes towards economics	1,289	2.802	0.373	2.833	1	5
Attitudes towards money	1,289	3.979	0.774	4	1	5
Financial planning	1,289	3.057	0.543	3.125	1	5
Financial autonomy	1,289	2.82	1.119	3.067	1	5
Impulse purchases	1,153	1.976	0.896	2	1	4
Any savings (1/0)	1,104	0.826	0.379	1	0	1

*Notes:*This table descriptive statistics for various outcomes relevant to economic decision-making. All outcomes are described in detail in Appendix C.

Table A2: Model fit statistics

	1-PL		2-PL		3-PL		4-PL	
Itemno.	$S-\chi^2$	p-val.	$S-\chi^2$	p-val.	$S-\chi^2$	p-val.	$S-\chi^2$	p-val.
1	39.167	0.000	16.280	0.061	10.900	0.207	6.308	0.504
2	67.419	0.000	35.543	0.000	20.356	0.009	16.124	0.024
3	23.878	0.008	18.320	0.032	13.557	0.094	12.879	0.075
4	70.039	0.000	15.733	0.073	5.616	0.690	5.573	0.590
5	26.669	0.003	7.496	0,586	6.926	0.545	6.838	0.446
6	12.006	0.285	11.262	0,258	6.680	0.572	4.278	0.747
7	31.745	0.000	9.761	0.370	6.004	0.647	6.521	0.480
8	9.025	0.530	4.617	0.866	4.753	0.784	4.267	0.749
9	15.763	0.107	10.224	0.333	10.636	0.223	10.649	0.155
10	21.464	0.018	12.487	0.187	8.454	0.390	5.994	0.540
11	44.313	0.000	38.902	0.000	7.699	0.463	7.024	0.426
12	152.926	0.000	34.656	0.000	13.445	0.097	15.186	0.034

Notes: This table shows chi-square statistics (Orlando and Thissen 2003) for unidimensional IRT models with one, two, three, and four parameters.

Table A3: Differential Item Functioning

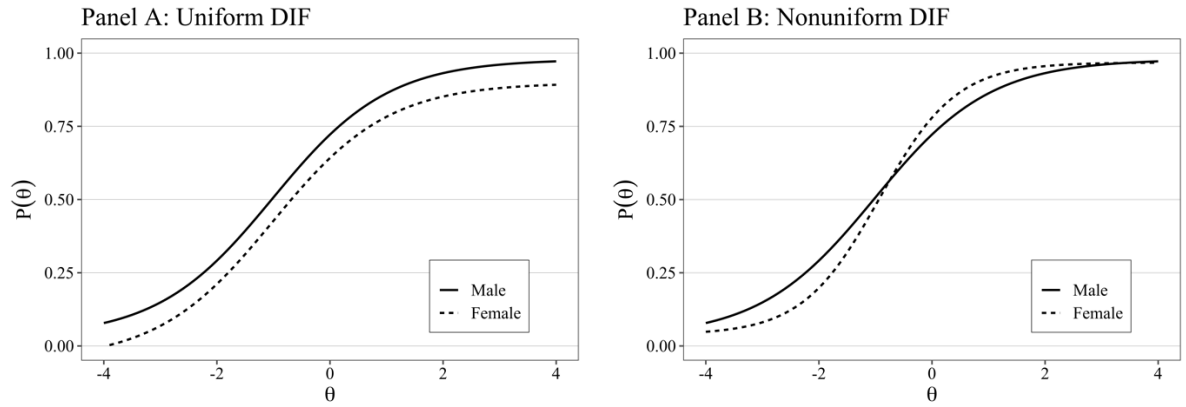
	Gender *(Focal group: Female)*		Books at home *(Focal group: <26 books at home)*		Native language *(Focal group: Non-natives)*	
Itemno.	*Δ-DIF* [*MHχ2*]	ETS	*Δ-DIF* [*MHχ2*]	ETS	*Δ-DIF* [*MHχ2*]	ETS
1	-1.160 [0.494]	B	-0.415 [0.176]	A	-0.608 [0.259]	A
2	0.488 [-0.208]	A	0.605 [-0.258]	A	0.242 [-0.103]	A
3	0.637 [-0.271]	A	0.452 [-0.192]	A	0.326 [-0.138]	A
4	0.380 [-0.162]	A	-0.666 [0.284]	A	-0.378 [0.161]	A
5	0.092 [-0.039]	A	-0.106 [0.045]	A	-0.054 [0.023]	A
6	1.248 [-0.531]	B	0.144 [-0.061]	A	-0.205 [0.087]	A
7	-0.333 [0.142]	A	0.066 [-0.028]	A	0.089 [-0.038]	A
8	0.261 [-0.111]	A	-0.132 [0.056]	A	0.159 [-0.068]	A
9	-0.030 [0.013]	A	0.208 [-0.089]	A	-0.172 [0.073]	A
10	-0.434 [0.185]	A	-0.054 [0.023]	A	0.128 [-0.054]	A
11	-0.167 [0.071]	A	-0.631 [0.269]	A	-0.270 [0.115]	A
12	-0.983 [0.418]	A	0.528 [-0.225]	A	0.743 [-0.316]	A

Notes: This table reports results from the analysis of differential item functioning along the split criteria gender, books at home and native language. Δ – DIF represents the Mantel-Haenszel delta difference as described in chapter 5.2 and MHχ2 its significance based on a χ2-distribution. Columns denoted as ETS report the classification according to the ETS scheme.

Appendix B: Robustness exercises

Nonuniform Differential Item Functioning. In chapter 5.2, we probe for uniform Differential Item Functioning, i.e., potential item bias is provided by a constant advantage of a subgroup over the other. However, the difference in the probability of endorsing an item between demographic subgroups may vary along the ability continuum (nonuniform DIF). For instance, male respondents may outperform females at low ability levels, whereas female respondents exhibit an advantage at high ability levels resulting in intersecting Item Characteristic Curves. Figure B1 illustrates an example for uniform (Panel A) and nonuniform DIF (Panel B) with respect to gender.

Figure B1: Examples of uniform and nonuniform Differential Item Functioning



Notes: This table shows examples of Item characteristic curves for male and female respondets graphically representing uniform DIF (left panel) as well as nonuniform DIF (right panel)

One widely used method for detecting nonuniform DIF is the logistic regression approach described in Swaminathan and Rogers (1990). Essentially, we regress latent trait estimates θ , the group membership (e.g., being female) and the interaction between these two components on the probability of solving an item correctly (y), formally expressed as $y = \beta_0 + \beta_1\theta + \beta_2 Group + \beta_3(\theta \times Group) + \epsilon$. If $\widehat{\beta_3}$ is significantly different from zero, the item exhibits nonuniform DIF. To categorize severity of DIF, we follow the scheme provided by Zumbo and Thomas (1997) where items are flagged as negligible DIF (category A) if the difference in the

pseudo R-squared between the aforementioned regression model and the reduced model without interaction term is below .13 ($\Delta R^2 \leq 0.13$). Items with $0.13 < \Delta R^2 \leq 0.26$ exhibit moderate DIF (category B) and items with $\Delta R^2 > 0.26$ are flagged as severe DIF (category C). Table A2 in Appendix A reports results with respect to gender, books at home and native language for the reduced 12-item scale. Overall, all items exhibit pseudo R^2 below 0.13 across all three demographic criteria therefore indicating no or only negligible nonuniform DIF.

Table B1: Nonuniform Differential Item Functioning (ZT-scheme)

	Gender *(Focal group: Female)*		Books at home *(Focal group: < 26 books at home)*		Native language *(Focal group: Non-natives)*	
Itemno.	$pseudo\ R^2$	Category	$pseudo\ R^2$	Category	$pseudo\ R^2$	Category
1	0.004	A	0.002	A	0.001	A
2	0.000	A	0.000	A	0.000	A
3	0.000	A	0.000	A	0.000	A
4	0.000	A	0.001	A	0.001	A
5	0.000	A	0.000	A	0.000	A
6	0.000	A	0.000	A	0.000	A
7	0.000	A	0.002	A	0.001	A
8	0.000	A	0.000	A	0.000	A
9	0.001	A	0.000	A	0.004	A
10	0.000	A	0.000	A	0.000	A
11	0.002	A	0.009	A	0.002	A
12	0.001	A	0.000	A	0.003	A

Notes: This table shows results for nonuniform Differential Item Functioning (NDIF) based on the logistic regression approach described in Swaminathan and Rogers (1990). Classification of DIF effects (category) is based on the method proposed by Zumbo and Thomas (1997). Socio-demographic variables as well as focal and references groups are selected as in Table 6.

Multidimensional Item Response Theory. To further prove the unidimensionality of the short scale, we test our one-dimensional IRT model against multidimensional IRT models. First, we test whether competence areas in the theoretical model described in section 2 represent dimensions of their own estimating a three-dimensional IRT model. Second, as we assumed missing values in competence items (item non-response) to be ignorable (missing at random), we estimate a two-dimensional IRT model that takes the propensity for item omissions into account. By following the approach in Pohl et al. (2014), we model - aside from person's ability

based on manifest observed responses - a latent missing propensity based on manifest missing responses.

Comparing the three-dimensional and the missing response model with the unidimensional model using Likelihood Ratio Tests reveals a significant better fit of the one-dimensional model to our data. The test statistic following a chi2-distribution was 113.57 ($p<0.01$) against the three-dimensional IRT model and 198.11 ($p<0.01$) against the two-dimensional model.

Appendix C: Item scales

(i) 12-item short scale of the Test of Economic Competence

1) (Itemno. 2 in the original scale)

Which statement about investing in shares is correct?

- Investing in shares is more secure than a savings account.
- Investment in shares can lead to losses.
- Investing in shares leads to constant income from interest.
- Investing in shares leads to constant income from dividends.

2) (Itemno. 3 in the original scale)

One day, the bakery "Empire Bread" mistakenly bakes more pumpkin-seed bread rolls than usually can be sold. Which measure would you recommend to the bakery on this day?

- Give away the remaining pumpkin-seed bread rolls.
- Increase the price of pumpkin-seed bread rolls on this day.
- Reduce the price of all of the bakery's products.
- Offer the pumpkin-seed bread rolls at a lower price

3) (Itemno. 6 in the original scale)

There is a regular flea market at school before the summer holiday. Emma in Class 8A owns the newest version of a popular video game she received from her aunt in Germany and which will only be released in the U.S. next year. She is considering selling it at the flea market. Which statement is correct?

- She would receive a comparatively high amount for the game this year
- She would receive a comparatively low amount for the game
- She would receive as much this year as she would receive next year
- She would not be able to sell the game this year
- She would not be able sell the game next year

4) (Itemno. 7 in the original scale)

An entrepreneur has set up a company manufacturing medical technology. When will the company start to generate profit? As soon as...

- the medical technology is sold in stores.
- income from the sales of the medical technology covers employees' monthly wages
- the company has crowded out all competing manufacturers of medical technology
- income from the sales of the medical technology covers monthly wages and the cost of renting manufacturing space.
- income from the sales of the medical technology is higher than all accrued costs.

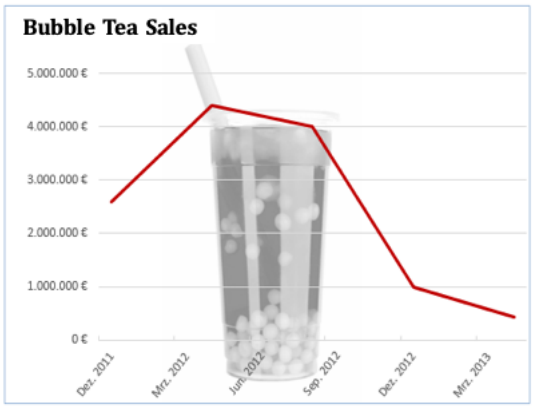
5) (Itemno. 8 in the original scale)

Michael had left school at age 16 and entered vocational training. How will Michael's income likely develop in comparison with the income of his former classmates, who continue their schooling and will later graduate from college?

- Michael's income will be higher than the income of his former classmates both now and in the future.
- Michael's income will be higher than the income of his former classmates now, but lower in the future.
- Michael's income will be lower than the income of his former classmates both now and in the future.

☐ Michael's income will be lower than the income of his former classmates now, but higher in the future.

6) (Itemno. 10 in the original scale)



This figure shows how the sales of bubble tea in Germany have developed in the course of 16 months. What can you conclude from the figure about sales of bubble tea?

☐ Bubble tea was banned in Germany since August 2012.
☐ Bubble tea continues to be sold profitably in Japan.
☐ Bubble tea is dangerous to health.
☐ Bubble tea was sold relatively little since August 2012.

7) (Itemno. 11 in the original scale)

In 1923, inflation in Germany was extremely high. With respect to the inflations' effect on retailers, which of the following statements is correct? Please select only one of the following answers:

☐ The inflation had no effect on retailers.
☐ They could set aside money for leaner times.
☐ They could pay their employees a higher salary.
☐ They no longer accepted cash as a means of payment.

8) (Itemno. 14 in the original scale)

Ms. Müller runs a dental surgery and makes €200 per hour. Today she is considering closing the surgery one hour earlier in order to mow the lawn at home. However, she could also hire a gardener for €50. Which statement is correct?

☐ She should mow the lawn herself in order to save the expense of the gardener.
☐ She should mow the lawn herself because she could do just as quickly.
☐ She should hire the gardener in order not to lose her income.
☐ It makes no difference because both cases involve one hour's work.

9) (Itemno. 16 in the original scale)

A sharp increase in the price of gasoline causes only a small decrease in the amount of gasoline sold in the short term. Why is this?

☐ Gasoline is a luxury good.
☐ The cost of gasoline makes up a large part of a household's expenditure.
☐ Gasoline cannot be easily replaced with something else.
☐ Taxes on gasoline are high.
☐ Vehicles do not need much gasoline nowadays.

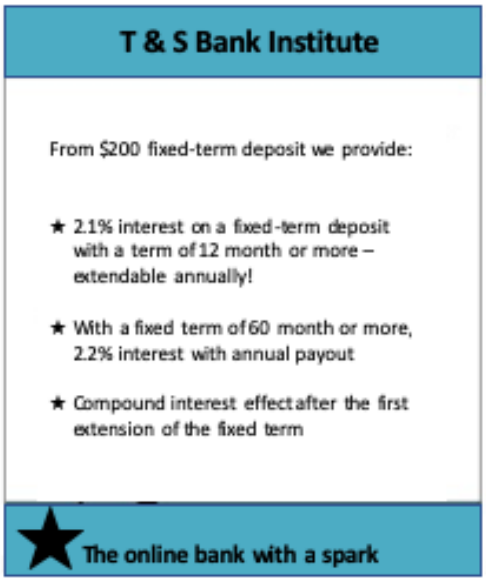
10) (Itemno. 20 in the original scale)

Two friends, Emil and Kadir, go to the bank. Emil borrows €1000 from the bank, Kadir deposits €1,000 into his savings account. After one year, Emil wants to pay back the money, and Kadir wants to withdraw his money.

- ☐ Emil has to pay back €1,000. Kadir receives €1,000.
- ☐ Emil has to pay back €1,000. Kadir receives more than €1,000.
- ☐ Emil has to pay back more than €1,000. Kadir receives €1,000.
- ☐ Emil has to pay back more than €1,000. Kadir receives more than €1,000; the amount is the same for both of them.
- ☐ Emil has to pay back more than €1,000. Kadir receives more than €1,000; Emil's amount is higher than Kadir's.

11) (Itemno. 25 in the original scale)

Finya got a gift of €2,000 from her grandparents for her sixteenth birthday. She would like to deposit the money into a bank-account. She finds these offers online:



T & S Bank Institute and BonusBank mention the effect of compound interest. What does this mean?

- ☐ The interest rate is highest in the first year.
- ☐ The interest rate rises from year to year.
- ☐ Interest will be paid the following year on interest paid out.
- ☐ The amount of money invested has an effect on the interest rate.
- ☐ The interest rate increases with annual credit.
- ☐ None of the above statements is correct.

12) (Itemno. 27 in the original scale)

Mr. Schneider receives a wage increase. He sees on his bank statement that, starting in January, he received almost exactly 1% more in his wage from his employer compared with January the previous year. The inflation rate for the previous year was 2%. Which statement is correct? Please select only one of the following answers:

- ☐ Mr. Schneider can afford more with his January wage than he could 12 months ago.
- ☐ Mr. Schneider can afford just as much with his Janu-ary wage as he could 12 months ago.
- ☐ Mr. Schneider can afford less with his January wage than he could 12 months ago.
- ☐ There is no connection

(ii) Measurement of constructs relevant to economic decision-making

Attitudes towards economics. To measure attitudes towards economics as a subject, we rely on a 12-item scale developed in Soper and Walstad (1983), adapted and translated in Oberrauch and Seeber (2021). For example, the scale asks participants the extent to which they disagree or agree with statements, such as "I follow economic news" or "It think it is important to have a good knowledge about economics".

Attitudes towards money. We measure this trait using the subdimension 'importance of money' from the Money Attitude Scale (MAS) originally developed by Yamauchi and Templer (1982). For instance, the scale asks students the extent to which they agree with statements, such as "Money is very important factor in the lives of all of us" or "Money is valueable".

Financial planning. To measure students' skills in financial planning, we also rely on a subscale of the MAS called "Time-Retention" describing behaviors that aim at the future and require thoughtful planning. The scale asks examinees the extent to which they agree with statements, such as "I budget my money very well" or "I keep track of my money".

Financial autonomy. We proxy socio-emotional skills related to financial decision-making using the Financial Autonomy Scale originally developed by and based on Noom et al. (2001) and employed in recent studies evaluating school financial or economic education interventions (Bruhn et al. 2016; Kaiser and Oberrauch 2021). In essence, the 15-item scale measures whether students feel "empowered, confident, and capable of making independent financial decisions" (Bruhn et al. 2016, p. 258). For example, it asks students the extent to which they agree with statements, such as "I make sure to get information on warranty periods" , "I feel prepared to talk to my parents about money matters", or "I suggest at home that we keep money aside for emergencies". Finally, we asked participants the extent to which they tend to spend new money to quickly on a scale from 1 ("No, not at all") to 4 ("Yes, totally") and whether they have any savings. Table C1 shows summary statistics for all six outcomes.

Financial autonomy scale (Bruhn et al. 2016)

Reflexive Autonomy:

- I like to think thoroughly before deciding to buy something
- I like to research prices whenever I buy something
- I make sure to get information on warranty periods
- I always try to obtain more information on product quality
- I pay attention to news about the economy as it may affect my family

Emotional Autonomy:

- I like to participate in family decision making when we buy something expensive for home
- I usually have a critical view of the way my friends deal with money
- I take part in domestic expense planning
- I try to advise my parents on money matters
- I feel prepared to talk to my parents about money matter.

Functional Autonomy:

- I always try to save some money to do things I really like
- I always like to negotiate prices when I buy
- I suggest at home that we keep money aside for emergencies
- I keep an eye on promotions and discounts
- I am willing to make sacrifices now to buy something important

Attitudes towards economics (Walstad and Soper 1983; Oberrauch and Seeber 2021)

- I enjoy reading articles about economic topics.
- Economics is easy for me to understand.
- Economics is dull.
- I'm interested in economics.
- I have nothing to say in economic matters.
- I like to talk about economics.
- I get bored in conversations about economics.
- I like to bring discussions to the topic economics
- I follow economic news.
- I wish I didn't have to learn economics.
- I learn a lot of interesting stuff when economics is discussed.
- I think it is important to have a good knowledge about economics.

Attitudes towards money (Tim and Leo 1997)

- Money is important
- Money is an important factor in the lives of all of us.
- Money is valuable
- I value money very highly.

Financial planning (Barry and Breuer 2012) (translated from German)

- I budget my money very well.
- I'm very careful with my money.
- I pride myself on my ability to save money.
- I keep track of my money.
- I regularly put money aside for the future.
- I often spend money even though I didn't plan to.
- I sometimes have to borrow money from others to make ends meet.
- I keep regular records of my income and expenses.

Appendix References

Bruhn, M., Leão, L. de S., Legovini, A., Marchetti, R., & Zia, B. (2016). The Impact of High School Financial Education: Evidence from a Large-Scale Evaluation in Brazil. *American Economic Journal: Applied Economics*, *8*(4), 256–295.

Kaiser, T. & Oberrauch, L. (2021). Economic education at the expense of indoctrination? Evidence from Germany, ZBW - Leibniz Information Centre for Economics, Kiel, Hamburg

Noom, M. J., Deković, M., & Meeus, W. (2001). Conceptual Analysis and Measurement of Adolescent Autonomy. *Journal of Youth and Adolescence 2001 30(5)*, *30*(5), 577–595.

Oberrauch, L. & Seeber, G. (2021). The impact of mandatory economic education on adolescents' attitudes. *Education Economics* 30 (2), 208-224

Orlando, M. & Thissen, D. (2003). Further investigation of the performance of S_X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement* 27 (4), 289–98.

Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests. Evaluating approaches accounting for missing responses in item response theory models. *Educational and Psychological Measurement,* 74 (3): 423–452.

Soper, J., & Walstad, W.B. (1983). On Measuring Economic Attitudes. *The Journal of Economic Education 14* (4), 4–17.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement,* 27 (4), 361–370.

Yamauchi, K. T., & Templer, D.I. (1982). The Development of a Money Attitude Scale." *Journal of Personality Assessment* 46 (5), 522–28.

Zumbo, B. D., & Thomas, D. R. (1997). A measure of effect size for a model-based approach for studying DIF (Tech. Rep.). Prince George, Canada:University of Northern British Columbia.
Zwinderman, A. H. (1991). A generalized Rasch model