J Fam Econ Iss (2014) 35:516–531
DOI 10.1007/s10834-013-9386-8

ORIGINAL PAPER

Evaluation of a Financial Literacy Test Using Classical Test Theory and Item Response Theory

Irina A. Kunovskaya · Brenda J. Cude · Natalia Alexeev

Published online: 12 January 2014

Abstract The purpose of this study was to evaluate the quality (item difficulty and discriminability, construct validity, and reliability) of a financial literacy test that has been used to measure financial literacy in multiple countries. The test was analyzed based on Classical Test Theory and Item Response Theory, specifically the Rasch model, with the goal of identifying areas of improvement. The findings from the two measurement frameworks were quite comparable and identified similar measurement problems. Steps to improve the financial literacy measure were recommended.

Keywords Measurement of financial literacy · Classical Test Theory · The Rasch model

Introduction

Interest in financial literacy has been on the rise internationally since the early 2000s, and a large number of governments from both developed and emerging countries have conducted financial literacy/capability[1] surveys designed to identify the level of their citizens' financial knowledge and financial behaviors. The Organisation for Economic Co-operation and Development (OECD) made a considerable impact with their landmark study (Organisation for Economic Co-operation and Development (OECD) 2005) that summarized the available results of surveys conducted in Australia, Japan, Korea, the United Kingdom, and the United States. Starting in 2008, the World Bank began diagnostic reviews of financial capability and consumer protection in several transitional economies and carried out seven[2] Financial Capability and Consumer Protection (FCCP) surveys at the national level during the period of 2008–2011 (World Bank Group 2011).

Most of the time, surveys are used as diagnostic tools to inform the design of national strategies to improve households' financial well-being and strengthen the

I. A. Kunovskaya (✉)
College of Family and Consumer Sciences, University of Georgia, 226 Hoke Smith Annex, Athens, GA 30602-4356, USA
e-mail: irkunovs@uga.edu

B. J. Cude
College of Family and Consumer Sciences, University of Georgia, 201 Housing Research Center, Athens, GA 30602-4356, USA
e-mail: bcude@uga.edu

N. Alexeev
Department of Mathematics, University of Georgia, 649A Boyd Graduate Studies Research Center, Athens, GA 30602-4356, USA
e-mail: nata@uga.edu

[1] The terms "financial literacy" and "financial capability" often have been used interchangeably in the literature; other times, they have been used to represent distinct but related concepts. Traditionally, financial literacy has been the preferred term in North America while financial capability is the term preferred in Great Britain (Social and Enterprise Development Innovations 2004, p. 5). The evolution of the title of US President's Advisory Council from Financial Literacy to Financial Capability in its second incarnation in 2010 may, however, signal a change in North America.

[2] The World Bank Financial Capability and Consumer Protection surveys were conducted twice in Russia (2008, 2009), and once in Azerbaijan (2009), Bulgaria (2010), Romania (2010), Bosnia and Herzegovina (2011), and the West Bank and Gaza (2011). An additional group of seven (Armenia, Colombia, Lebanon, Mexico, Nigeria, Turkey, and Uruguay) low- and middle-income countries was involved in the World Bank Financial Capability Survey conducted with the resources provided through the Russia Financial Literacy and Education Trust Fund (RTF) to evaluate the overall levels of financial capability in a population (Kempson et al. 2013).

financial sector. Surveys, focusing on financial knowledge, help to identify specific knowledge gaps and potential target populations for financial education (Perotti et al. 2013). The development of a highly flexible and applicable survey instrument to consistently and reliably measure financial capability at a national level and across countries has become a crucial task to inform policy as they develop financial education, promote access and use of financial services, and strengthen consumer protection mechanisms as well as social security systems (Atkinson 2008, 2012).

Advocates of financial literacy education highlight the importance of comparing the surveys' results across countries and developing international approaches to financial education, identifying good practices, and international priorities (Atkinson and Messy 2011; Holzman 2010). A common theme from the survey results suggests that whatever the country, financial understanding is low among consumers and particularly among the less educated, minorities, and those at the lower end of the income distribution (Kunovskaya and Cude 2012; OECD 2005; Perotti 2012).

A number of studies have, in fact, proposed frameworks and measurement instruments for financial capability surveys that allow comparisons across countries (Atkinson and Messy 2011; Kempson et al. 2012). The OECD recently piloted a questionnaire designed to capture international levels of financial literacy among adults in 14 developing and developed countries (Atkinson and Messy 2011) and among young people through PISA 2012 (OECD 2012). However, there is a growing concern as to how to assure that an instrument measures what it purports to measure—i.e., that financial literacy measures are reliable, valid, and fully comparable. The use of unreliable and invalid data may lead to unreliable and invalid results (Wright 1999). Multiple potential problems could affect the value of an instrument. For example, the distribution of the item difficulties could affect the distribution of the measures of respondents' financial capability expressed as raw scores (OECD 2009).

The purpose of this study was to evaluate the quality [item difficulty and discriminability, construct validity, and reliability (Embretson and Reise 2000)] of a financial literacy test that has been used to measure financial literacy in multiple countries. The evaluation used two currently popular statistical frameworks for addressing measurement problems: Classical Test Theory (CTT), a widely used traditional approach (Allen and Yen 1979) and Item Response Theory (IRT), an alternative and more sophisticated framework. The specific IRT model applied was the Rasch model, a one-parameter latent trait IRT model (Baker 2001; Bond and Fox 2001; Rasch 1960; Wright and Stone 1979). The data used were from the administration of a financial literacy test as part of the World Bank Financial Capability and Consumer Protection surveys in Azerbaijan (Azerbaijan Micro-finance Association 2010), Romania (Stanculescu 2010), and Russia (World Bank Group et al. 2011). In other words, this study empirically examined to what extent the item statistics from CTT and those from Rasch model vary across different countries for the same questions, identified measurement problems, and outlined steps to improve the financial literacy measure.

Literature Review

Designing Measures of Financial Literacy

Assessing financial literacy internationally through an objective test is a challenging task. The initial step involves mapping the concept of financial literacy and identifying the key concepts that need to be covered (Kempson 2009). Research shows there has been much debate but little consensus about the conceptualization of financial literacy. Remund (2010) reviewed 100 resources including research studies, expert opinions, and intervention programs and reported that the regulatory, research, and practitioners' communities have not yet reached consensus on what financial literacy is. Huston (2010) reviewed 71 studies (including some of the same ones that Remund cited) and found that almost one-half defined financial literacy as financial knowledge. Others defined financial literacy as broader than knowledge, but there were inconsistencies about the additional component. Some added money management skills (Jump$tart Coalition 2007; the President's Advisory Council on Financial Literacy 2008) while for others the additional component was the ability to make effective decisions (Lusardi and Tufano 2009; Mandell 2008; Noctor et al. 1992) or practical experience (Moore 2003). A review of the literature shows the current trend in this discussion: Researchers and practitioners are moving away from financial literacy, the original term, toward financial capability, the more recently coined term, as a more extended concept with a broader range of interests relating to consumer financial behavior (Yoong et al. 2013).

This breadth in the conceptual definitions of financial literacy[3] makes the second measurement step—development and application of an instrument to measure financial literacy (operationalization)—more challenging (Ardic et al. 2011). Without a valid and reliable measure, it is more difficult to identify mechanisms to use to affect financial literacy through financial education (Borden et al. 2008; Hung et al. 2009; Postmus et al. 2012).

[3] In this paper, financial literacy is synonymous with financial knowledge.

Several recent, but influential, analyses have attempted to address operationalisation challenges in measuring financial literacy/capability. The Organisation for Economic Co-operation and Development in its early report (2005) suggested the importance of a uniform questionnaire for different countries. It also highlighted the advantage of an objective test over a test with self-reported items to measure financial literacy across countries.

Lusardi and Mitchell (2011) described four guiding principles to design financial literacy measures, namely simplicity, relevance, brevity, and capacity to differentiate. In addition, other standards for questions were recommended: Questions should address the "ABC's" of financial literacy, relate to day-to-day financial decisions, differentiate between financial knowledge levels, and be limited in number.

Assessing understanding of fundamental financial literacy concepts with a limited number of questions has become a widely-accepted approach; specifically, internationally many research studies have used questions written by Lusardi and Mitchell (2005, 2007; van Rooij et al. 2011; Cole et al. 2009). In an early study, Lusardi and Mitchell (2005) tested knowledge of compound interest, inflation, and stock risk with one question each. They used similar but slightly different questions in a later study (Lusardi and Mitchell 2007), again using one question per concept; the questions were termed "Percentage Calculation," "Lottery Division," and "Compound Interest." van Rooij et al. (2011) measured the basic financial literacy of Dutch households with five of Lusardi and Mitchell's questions; they termed the concepts "Numeracy," "Compound Interest," "Inflation," "Time Value of Money," and "Money Illusion." As van Rooij et al. (2011) argued, these concepts lie at the basis of financial transactions and day-to-day financial decision making.

Recently an emergent critique of the above questions has appeared in the literature. Hastings et al. (2012) doubted the merits of Lusardi and Mitchell "Big Three" questions from the 2004 Household Retirement Study as an international standard to assess financial literacy. They noted the lack of evidence that this is the best and most comprehensive approach to measure financial literacy. Knoll and Houts (2012) argued that repeated use of a set of financial literacy questions does not provide any guarantee that the questions measure literacy in the most effective way. Carpena et al. (2012) suggested that this set of questions is "not necessarily comprehensive and may not be appropriate in many settings" (p. 8).

Kempson et al. (2012) suggested that questions designed to assess financial literacy should satisfy several criteria: work across countries/be adjustable for differences between cultures and economies, apply to the whole population, discriminate capable and less capable respondents, be unambiguous, permit use of alternative analytical methods, be income neutral, and avoid scales based on value judgments. An appropriate instrument also makes the results of the surveys "comparable across time and space" (Holzman 2010, p. 14). Huston (2010) cited literature (Kim and Mueller 1978) suggesting a minimum of three to five items per concept measured. Green and Frantom (2002) and Kimberlin and Winterstein (2008) identified some of the same criteria, summarizing the description of an appropriate measurement instrument as one that is simple and easy to use and that provides high-quality information—indicators of the instrument's reliability and validity.

Assessing Measures of Financial Literacy

Although there is extensive literature about how to design a financial literacy measurement, there has not been much work specifically evaluating these measures. However, the literature suggests two main test theories that can be used to assess the properties of a measurement instrument: Classical Test Theory (CTT) and Item Response Theory (IRT).

Classical Test Theory (CTT) has been the predominant measurement paradigm in the area of test analysis and introduced three concepts—test score (or observed score), true score, and error score. The major advantage of CTT is its relatively weak theoretical assumptions, which make CTT easy to apply in many testing situations (Fan 1998).The basic assumption in CTT is that the observed score is an estimate of the true score of the person with the band of unobservable measurement around it (Hambleton and Swaminathan 1985; Wiberg 2004). The focus of the "true score theory" is an analysis of the total test score, frequency of correct responses, frequency of responses, reliability of the test, and item-total correlation (Lucey 2005; Magno 2009). Although the CTT approach is widely known and used to examine the quality of a measurement instrument, it has shortcomings. Among these are that all of the statistics which describe items (i.e., item difficulty and item discriminating power) are dependent on the sample of individuals and this dependency reduces their utility (Hambleton 2000); all the items are treated as equal contributors to the total score (Allen and Yen 1979). In addition, the observed total score is item dependent; it is not an absolute characteristic of a test-taker and depends on the content of the test (Hambleton and Swaminathan 1985).

Some financial literacy researchers have used elements of CTT to obtain characteristics of financial literacy measurement instruments. For example, Lucey (2005) applied CTT to describe the reliability (consistency) and validity of the Jump$tart Coalition's 1997 and 2000 surveys. Both Atkinson et al. (2006) and Klapper, Lusardi, and Panos (2011) reported the percent who correctly answered each

question and reliability of their financial literacy measurements. However, researchers have not extensively evaluated financial literacy measures using CTT.

With advances in psychometrics since the beginning of the 1970s, Item Response Theory has witnessed an exponential growth (Wong et al. 2011). The basic assumption of the Item Response Theory is an independence of the latent ability of the test-taker on the content of the test (Lord and Novick 1968). The IRT allows analysis of the responses from a sample of individuals to a bank of items and assumes that the responses depend on nonmeasurable respondent characteristics (latent traits) and on item characteristics (Baker 2001). As a link function between the responses to the items and the latent trait, the logistic function often is used in IRT models (Hardouin 2007).

IRT may be used for several purposes. The most typical one is to estimate a person's latent trait. The other common uses are designing an instrument that measures specific traits, analyzing test items and identification of biased test items, and equating alternate forms of a test.

Within the general IRT framework, many models have been formulated and applied to real test data (Hambleton and Jones 1993). The three-parameter model is the most general one. Two-and one-parameter models are nested under the three-parameter model (Fan 1998). This paper focused on applying the Rasch model, which is known for being the one-parameter logistic model and using IRT for binary variables (Bond and Fox 2001; Rasch 1960, 1993). The Rasch model is an ability measurement technique adapted from educational and psychological studies. It has been widely used in education and health sciences for more than 40 years. A fundamental assumption of this approach is that the response probability of each subject to each item is a function of the ratio of a person's ability (person parameter) to the item difficulty (item parameter) (Rasch 1993). The probability of the correct response of a person j to an item i is given by

$$P_{ij}(x_{ij}=1)=\frac{\exp(\theta_j-\beta_i)}{1+\exp(\theta_j-\beta_i)} \quad (1)$$

where x_{ij} is the response of person j to item i, θ_j is the latent ability of person j, and β_i is the difficulty of item i.

The Rasch model was used in a recent study of the financial capability of mutual fund investors; the study examined investors' abilities and awareness of the terms and risks of mutual fund investments (Pellinen et al. 2011). The researchers used the model to identify the 40 most-qualified items to measure financial capability in the questionnaire. The findings suggested that data analysis based on the Rasch model gave valuable information about the test-takers' true knowledge.

Knoll and Houts (2012) applied IRT using the 2PL model to analyze financial literacy items from three national surveys (2007 American Life Panel, 2004 and 2008 Household Retirement Study, and 2009 National Financial Capability Study). They used the data to create a psychometrically sound 20-item financial knowledge scale. They also conducted validation exercises and determined that the scale was predictive of outcomes known to be associated with financial literacy. However, their goal was to create a scale, not to evaluate an existing scale, which is the goal of the current article.

The present study employed both Classical Test Theory and the Rasch model[4] (1960, 1993), which in theory differ significantly, to examine the quality of a six-item financial literacy measurement instrument with the expectation to obtain richer information about item statistics by using both approaches. Rather than using a single data source, the instrument was tested more stringently by using data from three different administrations of the instrument.

Data and Sample

The data used for this study were from surveys that were part of the World Bank's country-level diagnostic reviews of the legal and regulatory frameworks of consumer protection and financial capability in selected countries (World Bank Group et al. 2011). Data from three of the countries surveyed (Azerbaijan, Romania, and Russia) were used in this study. Although Russia is by far the larger country and is more urban than the other two, they are similar in that all three are rebuilding their economies after being part of the Soviet Union, or, in the case of Romania, post-war Soviet occupation. The literacy rates in the three countries also are similar (www.indexmundi.com).

In each of the three countries, the Financial Capability and Consumer Protection surveys were administered using a common methodology. The surveys asked about 50 similar but not always precisely the same questions to provide country-specific information regarding respondents' understanding of basic financial concepts, awareness of financial consumer rights, patterns of household financial management, use of financial services, and attitudes regarding financial markets. All data were collected between 2008 and 2010 via face-to-face interviews with respondents 18 years of age or older. Sample sizes were between 1,207 (in Azerbaijan) and 1,912 (in Romania). The Russian data were from 1,600 individuals in 2008. The samples were nationally representative of the countries'

[4] Few studies have attempted empirically to examine the similarities and differences in the parameters estimated using the two frameworks [read an extended review of the theories in Fan (1998)].

Table 1 Characteristics of the samples by countries: Azerbaijan, Romania, and Russia

Variables	Azerbaijan (n = 1,207)	Romania (n = 1,912)	Russia (n = 1,600)
Age (mean)	42.06	50.5	44.5
<29	22.0 %	14.7 %	24.7 %
30–49	49.3 %	34.3 %	37.0 %
50 and over	28.7 %	51.0 %	38.3 %
Gender			
Female	48.3 %	48.3 %	54.6 %
Male	51.7 %	51.7 %‘	45.4 %
Household size (mean)	4.4	2.6	3.0
Single	3.0 %	22.9 %	11.1 %
More than 1	97.0 %	77.1 %	88.9 %
Education			
High school or less	51.4 %	83.9 %	39.6 %
Vocational education or incomplete higher education	21.4 %	5.1 %	43.3 %
Higher education	27.2 %	11.0 %	17.1 %
Employment status			
In the labor force	50.1 %	43.5 %	56.4 %
Retired	17.2 %	39.0 %	24.3 %
Students	3.5 %	4.2 %	12.6 %
Other not in the labor force	29.2 %	13.3 %	6.7 %
Settlement type			
Urban	54.0 %	47.9 %	73.2 %
Rural	46.0 %	52.1 %	26.8 %

populations by gender, age, education, and areas of residence.

The summary statistics describing the samples are reported in Table 1. The gender distribution was relatively equal; just more than one-half were female in Russia. There were slightly more males than females in Azerbaijan and Romania. The mean age was 44.5 years old in Russia, 50.5 years old in Romania, and 42.0 years in Azerbaijan. The mean household size was higher in Azerbaijan (4.4) than in Romania (2.6) and Russia (3). Most respondents (97 % in Azerbaijan, 77.1 % in Romania, and 88.9 % in Russia) lived in households with two to four people, including themselves. Other characteristics of the participants are reported in Table 1.

Methods

Instrument

The focus of the analysis reported here is an evaluation of the quality of a popular six-item financial literacy test. Each of the six items (see Appendix) was adapted from questions used by Lusardi and Mitchell (2005, 2007), van Rooij et al. (2011), and Cole et al. (2009); their use followed previous researchers' approach to measure financial literacy with a limited number of questions.

Four of the six questions aimed to test consumers' financial numeracy skills and their knowledge of compound interest, inflation, and money illusion; the questions were changed slightly in each survey to adjust them to the national currencies of the three countries. The other two questions aimed to test consumers' understanding of what was termed "discounting" and an interest rate calculation. Each of the six items was a multiple choice question with one correct answer. Moreover, respondents could indicate they did not know.

Results

Classical Test Theory

To evaluate the quality of the financial literacy test based on the Classical Test Theory, item statistics were computed for the data for each country—specifically, the item difficulty (p value) and the items' ability to discriminate between low- and high-ability test takers (point biserial correlation). A one-way ANOVA was applied to test for difficulty differences among six items on financial literacy tests in Azerbaijan, Romania, and Russia and to rank items accordingly. Difficulty differed significantly across the six items in each country.[5] Tukey HSD tests were conducted on all possible pairwise contrasts and results are reported in Table 2 in addition to the p values, point biserial correlation coefficients and item difficulty ranks. The ability to reliably obtain a score on the financial literacy test in each country was estimated using the corresponding Cronbach's alphas (Cronbach 1958). Factor analysis was used to analyze the dimensionality of test. STATA 12 software was used to conduct these analyses.

First, the test was examined using the proportion who answered each item correctly.[6] Table 2 reports these p values. In general, tests are more reliable when there is a wide spread of p values across the entire .0–1.0 range with a concentration toward .5 (Varma 2010). While there was a spread of p values across the items in each country, the spread was not very wide nor was it concentrated toward .5 in the Romanian (range .239–.506) and the Russian results

[5] One-way ANOVA results for Azerbaijan, Romania, and Russia were correspondingly: F (5, 7236) = 84.32, p = .000; F (5, 11466) = 78.03, p = .000; and F (5, 9594) = 169.08, p = .000.

[6] The success rate of a particular pool of examinees on an item, well known as the p value of the item, actually is an inverse indicator of item difficulty, with a higher value indicating an easier item.

Table 2 Descriptive statistics for financial literacy tests by country: Classical Test Theory approach

Item	Item label	Azerbaijan (n = 1,207)			Romania (n = 1,912)			Russia (n = 1,600)		
		p value (SE) [confidence interval] (1)	*p* value rank (2)	r_{pb} (3)	*p* value (SE) [confidence interval] (4)	*p* value rank (5)	r_{pb} (6)	*p* value (SE) [confidence interval] (7)	*p* value rank (8)	r_{pb} (9)
1	Numeracy	$.467_{abcf}$ (.014) [.439, .496]	4	.425	$.408_{ac}$ (.011) [.563, .611]	2	.520	$.587_{ac}$ (.012) [.563, .611]	2	.557
2	Compound interest	$.462_{abc}$ (.014) [.434, .491]	4	.409	.240 (.010) [.220, .259]	4	.450	.395 (.012) [.371, .419]	4	.451
3	Inflation	$.458_{abc}$ (.014) [.430, .486]	4	.422	$.433_{ac}$ (.011) [.410, .455]	2	.540	$.544_{ac}$ (.013) [.519, .568]	2	.408
4	Money Illusion	.617 (.014) [.590, .645]	2	.268	$.321_{df}$ (.011) [.300, .342]	3	.389	.456 (.013) [.431, .480]	3	.302
5	Discount	.781 (.012) [.758, .805]	1	.369	.500 (.011) [.483, .528]	1	.615	.696 (.012) [.674, .719]	1	.525
6	Interest Rate	$.523_{f}$ (.014) [.495, .551]	3	.392	$.317_{df}$ (.011) [.296, .338]	3	.559	.253 (.011) [.231, .274]	5	.268
	Mean	.552		.268	.371		.389	.488		.268
	Min	.458		.425	.239		.615	.252		.557
	Max	.781			.506			.696		

Means that do not share subscripts differ at $p < .05$ in the Tukey honestly significant difference comparison. Numbers in brackets are 95 % confidence intervals of the means

(range .252–.696) [see Columns (4), and (7) in Table 2]. In addition, the spread of *p* values was inconsistent across countries (the spread ranged from .267 in Romania to .404 in Russia).

Based on Tukey HSD tests, items found to be not statistically different in difficulty were assigned the same rank. As a result, there are four levels of difficulties in Azerbaijan and Romania and five levels in the Russian results.

Comparison of the individual items' *p* values across countries demonstrates, that Item 5 "Discounting," had the highest *p* values in each country's test—.781 (Azerbaijan), .506 (Romania), and .696 (Russia). It also had the same ranking (1) when items were sorted by their *p* values (from highest to lowest). But there were noticeable fluctuations in the proportions who correctly answered the other items on the test as well as in their ranks. For example, more than one-half of the respondents correctly answered Item 4 "Money Illusion" (*p* value .617, rank 2) and Item 6 (*p* value .523, rank 3) in Azerbaijan. However, only one-third of the respondents answered Item 4 correctly in Romania (rank 3). Item 6 "Interest Rate" was the most difficult item in Russia (*p* value .252, rank 5) but was answered correctly by more than one-half (*p* value .523, rank 3) of the Azerbaijanian respondents.[7]

Each item also was evaluated based on its ability to discriminate between low- and high-ability test takers; the measure of this, the point biserial correlation coefficient (r_{pb}), also is reported in Table 2. This coefficient reflects the correlation between a dichotomous variable (item score 0 or 1) and a continuous variable (the total score on a test from 0 to the maximum number of test items) (Fan 1998; Varma 2010). A negative discrimination value means that

[7] The similarity of the questions used to assess financial literacy in the World Bank Financial Capability and Consumer Protection surveys to questions used by Lusardi and Mitchell (2005) in the US allows comparison. American respondents demonstrated a much higher level of understanding of Item 2 "Compound Interest" (*p* value .617) and Item 3 "Inflation" (*p* value .752) than respondents from Azerbaijan, Romania, or Russia.

respondents with the lowest total scores selected the correct answer more than respondents with the highest exam scores. Positive values of this coefficient are desirable; low values[8] of a point biserial correlation coefficient indicate an anomaly and the need for the questions to be reviewed. The r_{pb} reported in Table 2 was bias corrected (i.e., the contribution of an item score to the total score was removed before calculating the coefficient for the item).

Examination of Columns (3), (6), and (9) in Table 2 reveals that all of the correlation coefficients were positive. Most of the coefficients (excluding the coefficients for Item 4 "Money Illusion" in Azerbaijan and Item 6 "Interest Rate" in Russia) were .3 or higher. This means that for the majority of test takers, the pattern of high scorers getting the question right more than low scorers held true.

There were differences across countries: None of the coefficients for Azerbaijan were >.425 while four of the five coefficients for Romania were >.5. These results reflect that the size of a point biserial correlation is greatly impacted by the variance in the scores of the respondents and suggest the test is at best inconsistent in discriminating between low- and high-ability test takers.

Although there is no real relationship between the point biserial correlation and *p* value statistics, their comparison can help to indicate problems in a test's construction. For example, in Azerbaijan, Items 4 "Money Illusion" and 5 "Discounting" have "conflicting" statistics—a relatively low point biserial coefficient (.268, Column 3) and a high *p* value (.617, Column 1). The same problem is observed but in reverse for two items in the Romanian data. Item 2 "Compound Interest" had a lower *p* value (.239, Column 4) than point biserial correlation (.450, Column 6) as did Item 6 "Interest Rate" (.317, Column 4 vs. .559, Column 6). *p* Values indicate item difficulty while point biserial correlations indicate item quality. While the statistics cannot identify the reason for the problem, a best practice is a qualitative review of the items for their wording, content, and/or translation. It is often advisable (Varma 2010) that these items be removed from future testing.

The reliability of the test was examined using standardized and unstandardized values of Cronbach's alpha. Generally, composite scales are considered to be measuring the same thing if the alpha is between .7 and .8 (Cronbach 1958). The alpha values were fully acceptable only for Romania's financial literacy test (standardized $\alpha = .767$, unstandardized $\alpha = .768$). Corresponding Cronbach's alpha values were slightly lower than .7 for Russia (standardized $\alpha = .686$, unstandardized $\alpha = .687$) as well as for Azerbaijan (standardized $\alpha = .651$, unstandardized $\alpha = .650$).

In the final evaluation based on the CTT, the financial literacy test was assessed for "factorial validity" (Allen and Yen 1979) or unidimensionality. A single-factor solution was expected, meaning that items on the test measure only one dimension and share correlations that can be explained by one underlying factor.[9]

The factor analysis demonstrated that the test was unidimensional. The initial eigenvalues[10] greater than one showed that a single component explained around 49 % of the variance in the data in Azerbaijan and 63 % of the variance in the Romanian data. Even though the Kaiser (1960) criterion performed for the Russian data indicated that the second factor had an eigenvalue greater than one, a one component solution was chosen based on the amount of variance explained by the first factor (53 %).[11]

In sum, an evaluation of the six-item financial literacy test based on Classical Test Theory demonstrated that the instrument contained moderately high internal consistencies across countries, and each item made a unique contribution to the instrument and loaded on one factor. CTT-based statistics also identified several problems. The items did not exhibit the desired range in difficulties and their difficulty was inconsistent across countries. Additionally, the test was at best inconsistent in discriminating between low- and high-ability test takers. "Conflicting" *p* values and point biserial correlation coefficients suggested that four of the six items should be reexamined. The next section reports additional analysis based on the Rasch model which in theory overcomes the major weakness of CTT, that is, the sample dependency of item/person statistics.

The Rasch Model

Model Fit Assessment

The Rasch model uses a psychometric approach and requires assessing to what extent the model assumptions

[8] Cohen (1988) suggested interpreting the magnitude of r_{pb} in the following way: as negligible if r_{pb} is in the interval [0, .1), small if r_{pb} is in the interval [.1, .3), medium if r_{pb} is in the interval [.3, .5), and large if r_{pb} is in the interval [.5, 1).

[9] Before conducting the factor analysis, a polychoric correlation matrix was used (Kolenikov and Angeles 2004) to investigate correlations between variables. The average inter-item correlations were higher for the Romanian sample than for the Azerbaijian or Russian samples. Item 4 "Money Illusion" demonstrated the lowest inter-item correlation for the Azerbaijian sample and Item 6 "Interest rate" demonstrated the lowest inter-item correlation for the Russian sample. (Results are available from the first author.) Overall, the sets of correlations suggested that perhaps not every item on the test cohered well enough with the overall measure of financial literacy. In other words, at least some of the items did not adequately contribute to the measurement of the overall construct.

[10] The results are available from the first author.

[11] The amount of variance explained by the first factor in Russia (53 %) is greater than the amount of variance explained by the first factor in Azerbaijan (49 %).

Table 3 Financial literacy test: summary statistics of Rasch modeling for non-extreme persons by country

	Measure	SE	INFIT		OUTFIT	
			MNSQ	ZSTD	MNSQ	ZSTD
Azerbaijan (n = 1,207, non-extreme persons = 997)						
Mean	.24	1.00	1.0	.1	1.02	.1
SD	1.21	.12	.30	.7	.70	.8
Max	1.81	1.18	1.94	2.8	5.81	2.6
Min	−1.84	.88	.49	−1.3	.29	−1.1
Romania (n = 1,912, non-extreme persons = 1,212)						
Mean	−.02	.97	1.00	.1	.99	.1
SD	1.12	.11	.29	.8	.46	.8
Max	1.78	1.14	1.96	2.8	3.34	2.7
Min	−1.78	.87	.57	−1.7	.44	−1.6
Russia (n = 1,600, non-extreme persons = 1,281)						
Mean	.19	1.02	1.00	.1	1.00	.1
SD	1.24	.11	.39	.9	.80	.8
Max	1.94	1.18	2.41	3.1	6.48	2.8
Min	−1.93	.91	.54	−1.5	.32	−1.2

are valid for the given data. To analyze data based on the Rasch model, the WINSTEPS version 3.75.0 was employed.[12] It provided indices of the fit of the data to the model as well as the major psychometric characteristics of the instrument.

If the data fit the model, the mean square and standardized fit indices[13] should be close to 1.0 and .0, respectively (Green and Frantom 2002). The results in Table 3 suggest that these data do fit the model reasonably well as the mean square was 1.0 in each country (Column 3) and the standardized statistic was close to .0 (.1).

In the one-parameter Rasch model, estimates for person and item measures (called person ability and item difficulty, correspondingly) are independent. It allowed obtaining on a common logit scale two indices of separation and reliability (centralized to zero). The person (item) separation estimates measure the spread of persons (items) in standard error units and should have values >1.0 for a useful measurement instrument. Person separation helps to classify people, and item separation is used to verify the item hierarchy (Linacre 2009).

Person separation estimates for the test in each country were less than one (correspondingly, .55, .43, and .51), meaning that the test instrument was not sensitive enough

Table 4 Root mean square errors (RMSE), reliability, and separation of Rasch modeling for non-extreme persons and 6 non-extreme items by country

	Azerbaijan		Romania		Russia	
	Person	Item	Person	Item	Person	Item
RMSE	1.06	.08	1.03	.07	1.11	.07
Reliability	.23	.99	.15	.99	.20	1.00
Separation	.55	10.74	.43	10.73	.51	14.51

to distinguish between high and low performers (Table 4).[14] On the other hand, the separation values of the items were relatively high at 10.7 in two of the three countries (Azerbaijan and Romania), and even higher (14.5) in Russia, indicating a great spread of the items along the item difficulty hierarchy.

High reliability for persons assumes that persons with high measures actually do have higher measures than persons estimated with low measures. The estimated person reliability in all three countries was very low[15]—.23 for the 997 Azerbaijanian non-extreme respondents with an RMSE[16] of 1.06, .15 for the 1212 Romanian non-extreme respondents with an RMSE of 1.03, and .20 for the 1281 Russian non-extreme respondents with an RMSE of 1.11 (Table 4). Low values of person reliability might indicate a narrow persons' ability range, or may be related to the small number of items on the test. The errors (RMSE) in these estimates were relatively high and signaled that the data were not properly fit to the expected ability of test-takers and the level of test difficulty (Macayan and Ofalia 2011). The good item reliability indices and low errors (RMSE) for each country (.99 for Azerbaijan and Romania, 1 for Russia; the RMSEs were correspondingly .08, .07, and .07) reflected that sample sizes were large enough to locate the items on the latent variable or the variance in the test items' difficulty (Wright and Stone 1979).

To determine whether the data satisfied the fundamental assumption of the Rasch model of unidimensionality, Wright's Unidimensionality Index was calculated as the ratio of the person separation using model standard errors, and the person separation using real (misfit inflated) standard errors (Wright 1994). A value of this index above .9 indicates unidimensionality; .5 and below indicates multidimensionality. The total coefficients were 1.24 (Azerbaijan), 1.32 (Romania), and 1.35 (Russia), indicating unidimensionality. To further explore the unidimensionality of the test, a Rasch

[12] The one parameter Item Response Theory model (Parameter Logistic Model-1PL) is considered to be mathematically identical to Rasch modeling (Washburn 2009).

[13] The summary fit statistics at the item and person level are called infit (inlier-pattern-sensitive) and outfit (outlier-sensitive) statistics and reported as mean-square fit statistics (MNSQ) and as standardized statistics (ZSTD).

[14] Table 4 was constructed based on WINSTEP's output (available from the first author).

[15] Person reliability is conceptually similar to Cronbach's alpha or the traditional "test" reliability (Linacre 2009).

[16] RMSE is the Root Mean Square Standard Error and can be computed over the persons or over the items.

residual-based principal component analysis (PCA) was completed. The results from these PCAs vary from conventional factor analysis and are more indicative than definitive indicators (Linacre 2009). The software program WINSTEPS identifies secondary dimensions in the data by the decomposition of the observed residuals. Finding the component that explains the largest possible amount of variance in the residuals means that the residuals are not a random noise. That is, items in the test may be measuring some other shared dimension. Eigenvalues >2 for the first contrast typically indicate the presence of multiple dimensions and associations between data (Linacre 2009).

A PCA performed on the residuals demonstrated first contrast eigenvalues <2, ranging from 1.4 (Azerbaijan and Romania) to 1.6 (Russia).[17] It appeared that the unidimensionality assumption for the Rasch model held for all three data sets used in this study.

Item Fit and Difficulty Estimates

Additional statistics were calculated to evaluate the individual items on the test and to identify items that demonstrate unexpected behavior (Bond and Fox 2001) and inconsistent patterns of responses when people are confused or inattentive to task. The item fit statistics and the measure order of items on the test are displayed in Table 5.[18] Measure order (Column 1) is an estimate of the parameter or item difficulty. According to Linacre (2009, p. 200), the difficulty of an item is defined as "the point on the latent variable at which its high and low categories are equally probable." The reported logit values for the items' difficulty are arranged in Table 5 from the most to less difficult items. The difficulty spread of the items (Column 1) was between .65 and −1.71 in Azerbaijan, between 1.11 and −1.15 in Romania, and between 1.70 and −1.57 in Russia (Table 5). The relative difficulties of the items within each country were, for the most part, completely consistent with the relative difficulties identified in analyses based on the CTT (Table 2). Across the countries, item difficulties were similar between Russia and Romania. Items 2, 4, and 6 ("Compound Interest," "Money Illusion," and "Interest Rate") had average difficulty indices above average, and the remaining three items (1, 3, and 5; "Numeracy," "Inflation," and "Discount") had below average difficulty. In Azerbaijan, four items (1, 2, 3, and 6; "Numeracy," "Compound Interest," "Inflation," and "Interest Rate") had logits above zero, and only the logits for Items 4 and 5 ("Money Illusion," and "Discount") were above the neutral score.

Although the mean outfit MNSQ for all items was reported in the acceptable range,[19] careful examination of the statistics revealed that Item 4 "Money Illusion" behaved as a slightly misfitted item. It had the highest outfit MNSQ in Azerbaijan (1.31) and Romania (1.29) and the second highest in Russia (1.22). The high outfit ZSDT also indicated a problem with Item 4; they were outside the acceptable range in all three samples: 5.7 (Azerbaijan), 6.6 (Romania), and 5.3 (Russia). Given that outfit ZSDT is interpreted as a z-test, the recommended acceptable values are in the range of −2 to 2 (Linacre 2009). Item 6 "Interest Rate" also showed a slight misfit in two (Romania and Azerbaijan) of the three countries (Table 5). These results are consistent with the analysis based on CTT which also provided evidence of misfit for Items 4 "Money Illusion" and 6 "Interest Rate."

The narrower items' difficulty range in logits than the persons' ability range in logits was demonstrated graphically.

Items and Persons Distribution

Item-by-person maps demonstrate graphically both items and persons on the same logit scale[20] (Fig. 1.) The logit values are shown at the very left side of each map and range from −2 to 2 in each country. The distribution of persons is indicated by the "#'s" on the left side and the distribution of items is shown by the item numbers (FL1-FL6) on the right side of the vertical axis. The most able persons and the most difficult items are located at the top of the country's item-by-person map and the least able persons and the least difficult items are located at the bottom. It is not possible to differentiate people in these groups further using WINSTEPS. The three letters "M," "S," and "T" on the left and right sides of the vertical axis can be used to examine the difficulty spread of the items. On the left side, "M" indicates the mean score for the group of respondents; on the right side, "M" refers to the mean logit for the items. "S" and "T" indicate correspondingly one and two standard deviations above (below) mean. An item with average difficulty has a score of zero, meaning that it would be solved correctly with 50 % probability by a participant with average ability level.

The item-by-person maps illustrate some misalignment of persons and items in Azerbaijan and Russia, where the average person position was greater than the average item

[17] Results available from the first author.

[18] Tables 3, 4 and 5 were constructed based on WINSTEP's output.

[19] Wright (1994) described the acceptance range for infit or outfit MNSQ between .6 and 1.4.

[20] In Rasch modeling, a person's "ability" and an item's "difficulty" are measured on the same logit scale. An ideal scale would differentiate people between −3 logits and +3 logits (Washburn 2009).

Table 5 Financial literacy test: item statistics by country

Item label	Measure	SE	INFIT		OUTFIT		Point-measure correlation	Item
			MNSQ	ZSTD	MNSQ	ZSTD		
Azerbaijan (n = 1,207)								
"Inflation"	.65	.07	.93	−2.4	.89	−2.3	.63	3
"Compound interest"	.62	.07	.95	−1.7	.89	−2.4	.63	2
"Numeracy"	.59	.07	.92	−2.5	.89	−2.4	.63	1
"Interest rate"	.23	.07	.98	−.5	.98	−.5	.61	6
"Money illusion"	−.39	.08	1.20	5.3	1.31	5.7	.52	4
"Discount"	−1.71	.09	1.02	.4	1.13	1.3	.58	5
Mean	.00	.08	1.00	−.2	1.02	−.1		
SD	.84	.01	.09	2.7	.16	2.9		
Max	.65	.09	1.20	5.3	1.31	5.7		
Min	−1.71	.07	.92	−2.5	.89	−2.4		
Romania (n = 1,912)								
"Compound interest"	1.11	.07	.99	−.2	.97	−.5	.61	2
"Interest rate"	.44	.07	.86	−5.5	.82	−4.7	.70	6
"Money illusion"	.41	.07	1.20	7.0	1.29	6.6	.59	4
"Numeracy"	−.30	.07	1.02	.8	1.01	.3	.69	1
"Inflation"	0−.51	.07	1.00	.0	1.02	.5	.71	3
"Discount"	−1.15	.07	.93	−2.2	.81	−3.5	.76	5
Mean	.00	.07	1.0	0	.99	−.2		
SD	.74	.00	.11	3.8	.16	3.6		
Max	1.11	.07	1.20	7.0	1.29	6.6		
Min	−1.15	.07	.86	−5.5	.81	−4.7		
Russia (n = 1,600)								
"Interest rate"	1.7	.07	1.1	2.8	1.33	3.8	.50	6
"Compound interest"	.65	.07	.89	−4.0	.84	−3.7	.64	2
"Money illusion"	.24	.06	1.19	6.5	1.22	5.3	.55	4
"Inflation"	−.35	.07	1.05	1.6	1.09	1.9	.62	3
"Numeracy"	−.66	.07	.81	−6.2	.77	−4.9	.72	1
"Discount"	−1.57	.08	.93	−1.7	.76	−3.1	.70	5
Mean	.00	.07	1.00	−.2	1.00	−.1		
SD	1.03	.00	.13	4.3	.22	3.9		
Max	1.70	.08	1.19	6.5	1.33	5.3		
Min	−1.57	.06	.81	−6.2	.76	−4.9		

position, meaning that, on average, persons were "smarter" than the items (the M on the left side of the map was higher than the M on the right). All three item-by-person maps illustrate gaps in the coverage on the upper and on the lower ends of the continuum, signaling a need to add easier and harder items. According to the maps, there was some mistargeting between the distribution of persons and items on the graphs, demonstrated by numerous persons whose positions were below or above where the financial literacy items were measuring. Mistargeting could possibly cause imprecise estimates of the item or/and person parameters (Hagquist et al. 2009).

In sum, the results indicate that the data fit the one-parameter IRT model well. At the same time, the evaluation based on the Rasch model confirmed some criticisms of the test identified when the evaluation was based on the CTT. Specifically, the Rasch model evaluation indicated that the test would benefit from the addition of both easier and harder items. Also consistent with the evaluation guided by CTT, the Rasch analysis raised questions about the value of Items 4 "Money Illusion" and 6 "Interest Rate." The evaluation guided by the Rasch model suggested the test was unidimensional and provided additional support for retaining one factor for

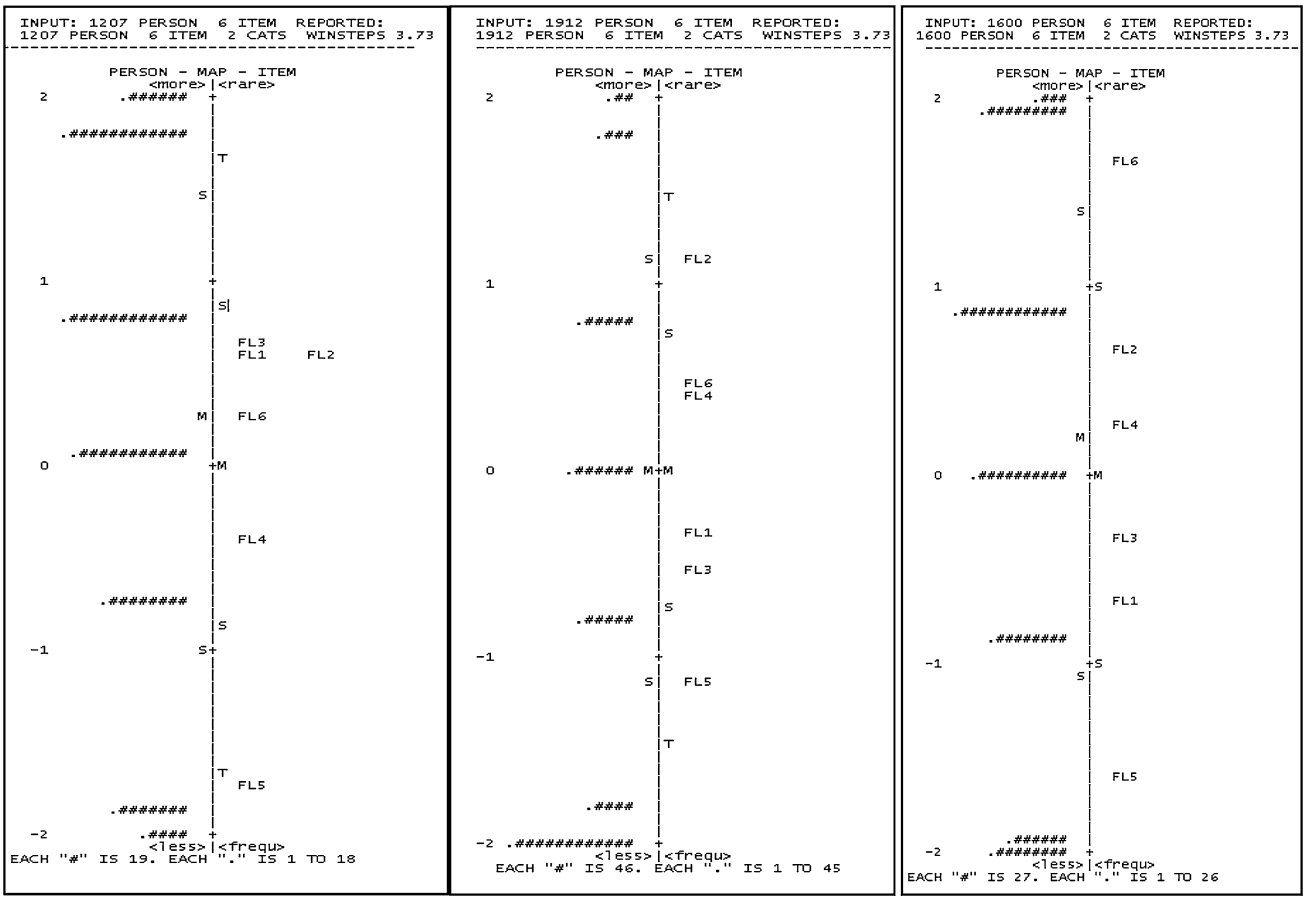

Fig. 1 Item-by person map: Azerbaijan, Romania, and Russia

the Russian data as suggested by the evaluation guided by CTT.

Discussion

This study aimed to examine the quality of a six-item financial literacy test which has been used to measure financial literacy in multiple countries. The properties used to assess the test's quality were examined using both Classical Test Theory and Item Response Theory (the Rasch model).

The evaluation using Classical Test Theory demonstrated the following properties of the financial literacy test. The internal reliability of the test was high in Romania (standardized $\alpha = .767$) and moderately high in Azerbaijan and Russia meaning the estimates of Cronbach's alpha were sample dependent. This confirms Magno's (2009) suggestion to estimate the internal consistency for each study even when one uses an instrument from past research.

Further, the *p* values of the items indicated the difficulty of the items had a more narrow range than desired—e.g., the test would benefit from the addition of both easier and more difficult questions. The point biserial correlation coefficients indicated the test was inconsistent at discriminating between low- and high-ability test takers; conflicting *p* values and point biserial coefficients raised questions about a few items. In addition, the difficulty of the items was inconsistent across countries. This may reflect the estimate's dependency on the sample attributable to CTT analysis (Magno 2009; Wiberg 2004).

The Rasch analysis of financial literacy test showed sufficiently good overall fit and high reliability measured by an item separation and reliability indexes. However, it also indicated problems with the person separation and reliability indexes, and identified two items (Item 4 "Money Illusion" and Item 6 "Interest Rate") as misfit.

Psychometric characteristics obtained by Rasch analysis point to the need for improvements and revisions in the financial test. First, low values of the person separation index indicated low variability of persons on the trait across all three countries. Person reliability indices (*real*) were rather low across the three samples as well—.23 (Azerbaijan), .15 (Romania), and .20 (Russia)—and were not sufficient to separate respondents into different performance levels. However, this result could be due to the relatively few items (six) on the test as well as there being

only three response options (plus "Don't Know") for five of the questions. The number of answer choices for the sixth, Item 5 "Discounting," was two in Russia and Romania but three in Azerbaijan (plus "don't know").

Collectively, the evaluation of the six-item financial literacy test demonstrated its acceptable reliability; however, it disclosed limitations in the instrument's ability to measure its targeted concept, despite its repeated use. The results suggest a number of ways in which the test should be improved.

- Additional items: As mentioned previously, good design principles call for three to five items per concept, not one as is the case for this test. Design of future evaluation instruments should start by identifying the concepts to be tested. Then, the researchers can write and pilot test multiple questions related to each concept. It also is important to consider how the relative number of response options per item influences outcomes (Washburn 2009).
- Addition of easier as well as harder questions: Good test design dictates the inclusion of easy as well as difficult questions. With a greater number of questions in the test, the range of question difficulty can be greater.
- Reexamination of at least four of the items in the current test: If Items 2, 4, 5, and/or 6 are to be included in future tests, their wording should be reexamined. For example, Item 4 "Money Illusion" could be interpreted differently by different respondents. Clearly, the correct "textbook" answer is that proportional increases in money income and consumer prices result in purchasing power actually remaining the same. However in reality, one may have doubts about whether "the same" actually means the same, knowing based on personal experience that not all prices for goods and services change in the same proportion and at the same time.
- Reexamination of the items also should include asking whether the questions that include numbers are actually testing financial literacy, numeracy, or some combination of both. Also, when tests are administered in multiple languages, the translation is key to ensure that the translated question and response choices still address the same construct.

The CTT and the Rasch model analyses provided the same or very similar results. The limited numbers of items on the test may explain the lack of superior performance by Rasch framework in our study. It is possible that other IRT models, which include parameters for differences in item discrimination (2-PL), and guessing (3-PL), would provide more plausible and richer information about the validity of the financial literacy test. Overall, the Rasch model appeared to be useful for a rigorous examination of the financial literacy measurement instrument confirmed the results of Classical Test Theory and disclosed measurement problems, but at the cost of greater model complexity.

However, the research is not without limitations, primarily in the data used to apply the models. There may be unique characteristics of the data that the evaluation could not identify, such as differences in interpretation of the questions, due perhaps to differences in the cultural knowledge shared by persons from different countries. It is possible that these unobserved differences might explain some of the results reported here. The study could not interpret the surveys' validity by reviewing prior financial literacy measures across countries, because of the lack of similar surveys.

Conclusion

As previously stated, both Classical Test Theory and the Rasch model (a one-parameter model from the family of Item Response Theory models) were useful to evaluate a six-item financial literacy test. They provided measures of item difficulty and discriminability, construct validity, and reliability. The research disclosed a variety of measurement problems and supports an emerging consensus that the field currently lacks a comprehensive measure of financial literacy, which enables policymakers to identify the needs of the population in financial literacy education and to set benchmarks for national strategies. The number of items on the test used in the World Bank Financial Capability and Consumer Protection Surveys in Azerbaijan, Romania, and Russia was insufficient to cover all content of financial literacy concept. Specifically, to strength the validity before future use, we suggest researchers write more items (both easier and more difficult items), review the content of most of the existing items, and, perhaps most importantly, revisit the construct definition of the instrument to adapt it to the relevant cultural environment (Hambleton 2005). Having in each country an established measure as a standard for the surveys is also important.

Each of these actions is expected to improve the reliability and validity of the instrument as a tool to measure financial literacy across different countries and throughout diverse populations and to improve the items' difficulty and discriminability.

Appendix

The financial literacy test: the country-specific versions of Question 1 are given, Questions 2–6 are those in the Russian survey but similar modifications were made as needed to account for the different currencies in each country.

Item	Concepts	Question
1	"Numeracy" (Russia)[a]	Let's assume that you deposited 100,000 rubles in a bank account for 2 years at 8 % interest rate. How much money will you have in your account in 2 years if you do not withdraw from or add to this account any money?
		1. More than 108,000 rubles
		2. Exactly 108,000 rubles
		3. <108,000 rubles
		99. I cannot come up with even a rough number
	'Numeracy" (Romania)	Let's assume that you deposited 10,000 lei in a bank account for 2 years at 8 % interest rate. How much money will you have in your account in 2 years if you do not withdraw from or add to this account any money?
		1. More than 18,000 lei
		2. Exactly 18,000 lei
		3. <18,000 lei
		99. Don't know
	'Numeracy" (Azerbaijan)	Let's assume that you deposited 1,000 manats in a bank account for 1 year at 12 % interest rate. How much money will you have in your account in 1 year if you do not withdraw from or add to this account any money?
		1. More than 1,120 manats
		2. Exactly 1,120 manats
		3. <1,120 manats
		99. I cannot come up with even a rough number
2	"Compound interest"	Let's assume that you deposited 100,000 rubles in a bank account for 5 years at 10 % interest rate. The interest will be earned at the end of each year and will be added to the principal. How much money will you have in your account in 5 years if you do not withdraw either the principal or the interest?
		1. More than 150,000 rubles
		2. Exactly 150,000 rubles
		3. <150,000 rubles
		99. I cannot estimate the amount even roughly
3	"Inflation"	Imagine, than you deposited the money in a bank account at 8 % interest rate, while the annual inflation rate was 10 %. Do you think the money from your account can buy more or less, or the same amount of goods and services on average now as a year ago?
		1. More than a year ago
		2. The same
		3. Less than a year ago
		99. I cannot estimate it even roughly
4	"Money illusion"	Let's assume that in 2010 your income is twice as much now, and the consumer prices also grow twofold. Do you think that in 2010 you will be able to buy more, less, or the same amount of goods and services as today?
		1. More than today
		2. Exactly the same
		3. Less than today
		99. I cannot estimate it even roughly
5	"Discount"	Let's assume that you saw a TV-set of the same model on sale in two different shops. The initial retail price of it was 10,000 rubles. One shop offered a discount of 1,500 rubles, while the other one offered a 10 % discount. Which one is a better bargain—a discount of 1,500 rubles or 10 %?
		1. A discount of 1,500 rubles
		2. A 10 % discount
		99. I cannot estimate it even roughly
6	"Interest rate"	Let's assume that you took a bank credit of 10,000 rubles to be paid back during a year in equal monthly payments. The credit charge is 600 rubles. Give a rough estimate of the annual interest rate on your credit.

continued

Item	Concepts	Question
		The interest rate is about: 1. 3 % **2. 6 %** 3. 9 % 4. 12 % 99. I cannot estimate it even roughly

This question in the earlier studies was referred to as "Compound Interest" question (Lusardi and Mitchell 2005, 2007). In this study we kept the name "Numeracy" for Item 1 since Item 2 clearly tested the concept of compound interest. This item had slightly different wording in Russia compare to Romania and Azerbaijan resulting in different correct options. There were no noticeable differences across Items 2, 3, 4, and 6. Item 5 in Azerbaijan sample had 3 response options instead of 4 as it was in Romania and Russia

References

Allen, M., & Yen, W. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

Ardic, O. P., Ibrahim, J. A., & Mylenko, N. (2011). *Consumer protection laws and regulations in deposit and loan services: A cross-country analysis with a new data set* (Policy Research Working Paper 5536). The World Bank. doi:10.1596/1813-9450-5536.

Atkinson, A. (2008). *Evidence of impact: An overview of financial education evaluations*. University of Bristol: Personal Finance Research Centre. Retrieved January 12, 2012, from www.fsa.gov.uk/pubs/consumer-research/crpr68.pdf.

Atkinson, A. (2012). *OECD/INFE tools for cross-country surveys of financial literacy*. Colombia-OECD-WB Conference on Financial Education. Cartagena, Colombia: OECD. Retrieved November 4, 2012, from http://www.oecd.org/daf/financialmarketsinsuranceandpensions/financialeducation/oecd-colombiainternationalconferenceonfinancialeducation.htm.

Atkinson, A., McKay, S., Kempson, E., & Collard, S. (2006). *Levels of financial capability in the UK: Results of a baseline survey*. Financial Services Authority. Retrieved January 12, 2012, from www.fsa.gov.uk/pubs/consumer-research/crpr47.pdf.

Atkinson, A., & Messy, F. (2011). Assessing financial literacy in 12 countries: An OECD pilot exercise. *Journal of Pension Economics and Finance, 110*, 657–665. doi:10.1017/S1474747211000539.

Azerbaijan Micro-finance Association. (2010). *Final report: Results of the financial literacy survey*. Baku, Azerbaijan: Author. Retrieved November 20, 2011, from http://siteresources.worldbank.org/INTECAREGTOPPRVSECDEV/Resources/Financial_Literacy_Survey_Report_EN.pdf.

Baker, F. (2001). *The basics of item response theory. ERIC clearinghouse on assessment and evaluation*. College Park: University of Maryland.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. London: Lawrence Erlbaum Associates.

Borden, L. M., Lee, S., Serido, J., & Collins, D. (2008). Changing college students' financial knowledge attitudes and behavior through seminar participation. *Journal of Family and Economic Issues, 29*, 23–40. doi:10.1007/s10834-007-9087-2.

Carpena, F., Cole, S., Shapiro, J., & Zia, B. (2012). *Unpacking the causal chain of financial literacy* (Policy Research Working Paper 5798). Washington, DC: The World Bank.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cole, S., Sampson, T., & Zia, B. (2009). *Financial literacy, financial decisions, and the demand for financial services: Evidence from India and Indonesia* (Working Paper 09-117). Harvard Business School, Harvard University. Retrieved January 12, 2012, from http://ssrn.com/abstract=1374078.

Cronbach, L. J. (1958). Proposals leading to analytic treatment of social perception scores. In R. Tagiuri & L. Petrullo (Eds.), *Person perception and interpersonal behavior* (pp. 353–379). Stanford, CA: Stanford University Press.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*(3), 357–381. doi:10.1177/0013164498058003001.

Green, K., & Frantom, C. (2002, November). *Survey development and validity with Rasch model*. A paper presented at the International Conference on Questionnaire Development, Evaluation, and Testing. Charleston, SC. Retrieved January 12, 2012, from http://www.jpsm.umd.edu/qdet/final_pdf_papers/green.pdf.

Hagquist, C., Bruce, M., & Gustavson, P. (2009). Using the Rasch model in nursing research: An introduction and illustrative example. *Journal of Nursing Studies, 46*, 380–393. doi:10.1016/j.ijnurstu.2008.10.007.

Hambleton, R. K. (2000). Emergence of item response modeling in instrument development and data analysis. *Medical Care, 38*(Suppl II), II-60–II-65. doi:10.1097/00005650-200009002-00009.

Hambleton, R. K. (2005). Issues, designs and technical guidelines for adapting tests into multiple languages and cultures. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberg (Eds.), *Adapting educational and psychological tests to cross-cultural assessment* (pp. 3–38). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff Publishing.

Hardouin, J.-B. (2007). Rasch analysis: Estimation and tests with Rasch test. *Stata Journal, 7*(1), 22–44.

Hastings, J. S., Madrian, B. C., & Skimmyhorn, W. L. (2012). *Financial literacy, financial education and economic outcomes* (NBER Working Paper 18412). Retrieved December 4, 2011, from http://www.nber.org/papers/w18412.

Holzman, R. (2010). *Bringing financial literacy education to low and middle income countries. The need to review, adjust, and extend current wisdom* (Discussion Paper No. 5114). Bonn: The Institute for the Study of Labor (IZA). Retrieved December 4, 2011, from http://ssrn.com/abstract=1663134.

Hung, A. A., Parker, A. M., & Yoong, J. K. (2009). *Defining and measuring financial literacy* (RAND Corporation Working Paper 708). Retrieved December 4, 2011, from http://www.rand.org/pubs/working_papers.

Huston, S. (2010). Measuring financial literacy: A proposed approach. *Journal of Consumer Affairs, 45*(2), 296–316. doi:10.1111/j.1745-6606.2010.01170.x.

Jump$tart Coalition. (2007). *National standards in K-12 personal finance education.* Retrieved December 4, 2011, from http://www.jumpstart.org/national-standards.html.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141–151. doi:10.1177/001316446002000116.

Kempson, E. (2009). *Framework for the development of financial literacy baseline surveys: A first international comparative analysis*. Paris, France: OECD. Retrieved November 2, 2012, from www.oecd.org/daf/fin/wp.

Kempson, E., Scott, V., & Perotti, V. (2012). *Measuring financial capability in a low- and middle-income setting.* Colombia-OECD-WB Conference on Financial Education. Cartagena, Colombia: OECD. Retrieved November 2, 2012, from http://www.oecd.org/daf/financialmarketsinsuranceandpensions/financialeducation/oecd-colombiainternationalconferenceonfinancialeducation.htm.

Kim, J.-O., & Mueller, C. W. (1978). *Factor analysis: Statistical methods and practical issues*. Newbury Park: SAGE.

Kimberlin, C., & Winterstein, A. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy*, *65*(23), 2276–2284.

Klapper, L. F., Lusardi, A., & Panos, G. A. (2011). *Financial literacy and the financial crisis: Evidence from Russia.* Retrieved December 4, 2011, from http://ssrn.com/abstract=1786826 or http://dx.doi.org/10.2139/ssrn.1786826.

Knoll, M., & Houts, C. (2012). The financial knowledge scale: An application of Item Response Theory to the assessment of financial literacy. *Journal of Consumer Affairs, 46*(3), 381–410.

Kolenikov, S., & Angeles, G. (2004). *The use of discrete data in PCA: Theory, simulations, and applications to socioeconomic indices* (Working Paper 04-85). MEASURE/Evaluation Project. University of North Carolina, Chapel Hill: Carolina Population Center. Retrieved December 4, 2011, from http://www.cpc.unc.edu/measure/publications/wp-04-85.

Kunovskaya, I., & Cude, B. (2012, April). Comparative analysis of money management as a component of financial capability in transitional economies. In *Proceedings of the 2012 American council on consumer interests annual conference*. Retrieved January 4, 2013, from http://www.consumerinterests.org/pdffiles/2012-conference/2012-9%20Comparative%20Analysis%20of%20Money%20Management.pdf.

Linacre, J. M. (2009). *A user's guide to WINSTEPS: MINISTEPS Rasch Model computer programs* (Program Manual 3.69.0). Retrieved from http://www.winsteps.com/index.htm.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lucey, T. (2005). Assessing the reliability and validity of the Jump$tart survey of financial literacy. *Journal of Family and Economic Issues, 26*(2), 283–284. doi:10.1007/s10834-005-3526-8.

Lusardi, A., & Mitchell, O. S. (2005). *Financial literacy and planning: Implications for retirement wellbeing* (Michigan Retirement Research Center Research Paper No. WP 2005-108). doi: 10.2139/ssrn.881847.

Lusardi, A., & Mitchell, O. S. (2007). Baby boomer retirement security: The roles of planning, financial literacy, and housing wealth. *Journal of Monetary Economics, 54*(1), 205–224. doi:10.1016/j.jmoneco.2006.12.001.

Lusardi, A., & Mitchell, O. S. (2011). *Financial literacy around the world: Introduction and overview* (CeRP Working Paper 106/11). Retrieved January 20, 2012, from http://www.cerp.carloalberto.org/en/publications.

Lusardi, A., & Tufano, P. (2009). *Debt literacy, financial experiences, and overindebtedness* (NBER Working Paper 14808). Retrieved February 10, 2011, from http://www.nber.org/papers.

Macayan, J., & Ofalia, B. (2011). Determining experts and novices in college algebra: A psychometric test development and analysis using the Rasch model (1PL-IRT). *Educational Measurement and Evaluation Review, 2*, 62–76.

Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *International Journal of Educational and Psychological Assessment, 1*(1), 1–11. Retrieved December 4, 2011, from http://ssrn.com/abstract=1426043.

Mandell, L. (2008). Financial literacy of high school students. In J. J. Xiao (Ed.), *Handbook of consumer finance research* (pp. 163–183). New York, NY: Springer. doi:10.1007/978-0-387-75734-6_10.

Moore, D. (2003). *Survey of financial literacy in Washington State: Knowledge, behavior, attitudes, and experiences* (Technical Report n. 03-39). Washington State University, Social and Economic Sciences Research Center. Retrieved February 14, 2009, from www.dfi.wa.gov/news/finlitsurvey.pdf.

Noctor, M., Stoney, S., & Stradling, S. (1992). *Financial literacy: A discussion of concepts and competences of financial literacy and opportunities for its introduction into young people's learning (Report prepared for the National Westminster Bank)*. London: National Foundation for Education Research.

Organisation for Economic Co-operation and Development (OECD). (2005). *Improving financial literacy: Analysis of issues and policies.* Paris, France: OECD. Retrieved January 20, 2012, from http://www.oecd.org/document/2/0,3746,en_2649_15251491_35802524_1_1_1_1,00.html.

Organisation for Economic Co-operation and Development (OECD). (2009). *PISA 2009 results: What students know and can do.* Paris, France: OECD. Retrieved November 5, 2012, from www.oecd.org/edu/pisa/2009.

Organisation for Economic Co-operation and Development (OECD). (2012). *PISA 2012 financial literacy framework*. Paris, France: OECD. Retrieved November 5, 2012, from www.oecd.org/dataoecd/8/43/46962580.pdf.

Pellinen, A., Törmäkangas, K., Uusitalo, O., & Raijas, A. (2011). Measuring the financial capability of investors: A case of the customers of mutual funds in Finland. *International Journal of Bank Marketing, 29*(2), 107–133. doi:10.1108/02652321111107611.

Perotti, V. (2012, October). *Review of existing financial capability and literacy measurement instruments.* Colombia-OECD-WB Conference on Financial Education. Cartagena, Colombia: OECD. Retrieved November 5, 2012, from http://www.oecd.org/daf/financialmarketsinsuranceandpensions/financialeducation/oecd-colombiainternationalconferenceonfinancialeducation.htm.

Perotti, V., Siegfried, Z., Larassi, G., & Bolaji-Adio, A. (2013). *Making sense of financial capability surveys around the world.* Washington: International Bank for Reconstruction and Development. Retrieved June 12, 2013, from http://responsiblefinance.worldbank.org/~/media/GIAWB/FL/Documents/Misc/Financial-Capability-Review.pdf.

Postmus, J., Plummer, S., McMahon, S., & Zurlo, K. (2012). Financial literacy: Building economic empowerment with survivors of violence. *Journal of Family and Economic Issues, 34*, 275–284. doi:10.1007/s10834-012-9330-3.

President's Advisory Council on Financial Literacy. (2008). *2008 Annual report to the President*. Washington, DC: The Department of the Treasury. Retrieved January 20, 2012, from http://www.treasury.gov/resource-center/financial-education/Documents/PACFL_ANNUAL_REPORT_1-16-09.pdf.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Rasch, G. (1993). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago.

Remund, D. L. (2010). Financial literacy explicated: The case for a clearer definition in an increasingly complex economy. *Journal of Consumer Affairs, 44*(2), 276–295. doi:10.1111/j.1745-6606.2010.01169.x.

Social and Enterprise Development Innovations (SEDI). (2004, July). *Financial capability and poverty: New approaches for addressing poverty and exclusion* (Discussion Paper). Toronto, Canada: Author. Retrieved January 20, 2012, from http://www.sedi.org/DataRegV2-unified/sedi-Publications/PRIEnglishAug2004.pdf.

Stănculescu, M. (2010, June). *Analysis of the financial literacy survey in Romania and recommendations.* Bucharest, Romania: Institute for the Study of the Quality of Life. Retrieved January 20, 2012, from http://siteresources.worldbank.org/INTECAREGTOPPRVSECDEV/Resources/Romania_Financial_Literacy_June_2010.pdf.

van Rooij, M., Lusardi, A., & Alessie, R. (2011). Financial literacy and stock market participation. *Journal of Financial Economics, 101*(2), 449–472. doi:10.1016/j.jfineco.2011.03.006.

Varma, S. (2010). *Preliminary item statistics using point-biserial correlation and p-values*. Educational Data Systems, Inc. Retrieved January 20, 2012, from http://www.eddata.com/resources/publications.

Washburn, I. J. (2009). *Rasch modeling in family studies: Modification of the relationship assessment scale*. (Master's thesis) Retrieved January 20, 2012, from http://ir.library.oregonstate.edu/xmlui/bitstream/handle/1957/11182/Take%20the%20Relationship%20Assessment%20Scale%20and%20redesign%20it%20using%20Rasch%20Modeling.pdf?sequence=1.

Wiberg, M. (2004). *Classical test theory vs. item response theory: An evaluation of the theory test in the Swedish driving license-test*. Umea: Umea Universiteit. Retrieved January 20, 2012, from http://www.jus.umu.se/digitalAssets/59/59529_em-no-50.pdf.

Wong, H. M., McGrath, C. P. J., & King, N. M. (2011). Rasch validation of the early childhood oral health impact scale. *Community Dentistry and Oral Epidemiology, 39*, 449–557. doi:10.1111/j.1600-0528.2011.00614.x.

World Bank Group, Financial and Private Sector Development Vice Presidency, FPDFS—Financial Systems Policy Unit. (2011). *Good practices for financial consumer protection* (Consultative Draft). Washington, DC: Author. Retrieved January 20, 2012, from http://siteresources.worldbank.org/EXTFINANCIALSECTOR/Resources/Good_Practices_Financial_CP.pdf.

Wright, B. D. (1994). Unidimensionality coefficient. *Rasch Measurement Transactions, 8*(3), 385.

Wright, B. D. (1999). Fundamental measurement for psychology. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 65–104). Mahwah, NJ: Lawrence Erlbaum Associates.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA Press.

Yoong, J., Mihaly, K., Bauhoff, S., & Rabinovich, L. (2013). *A toolkit for the evaluation of financial capability programs in low- and middle income countries*. Washington: International Bank for Reconstruction and Development. Retrieved June 12, 2013, from http://documents.worldbank.org/curated/en/2013/01/18054632/toolkit-evaluation-financial-capability-programs-low-middle-income-countries.

Irina A. Kunovskaya is a Director of Research at the Financial Literacy Group in Washington, DC and Evaluation Specialist at the University of Georgia College of Family and Consumer Sciences. Dr. Kunovskaya's research focuses on household finance, with a focus on savings, banking, financial capability, and inclusion.

Brenda J. Cude is a professor in the Department of Housing and Consumer Economics at the University of Georgia. Her current research focuses on financial literacy, specifically among the college student population. In addition, she is the Director of the University of Georgia Center on Economic Education.

Natalia Alexeev is a lecturer at the Department of Mathematics at the University of Georgia. Her BS and MS are in Mathematics from the Moscow State University, Russia, and her PhD is in Mathematics Education from the University of Georgia. As a postdoctoral associate at the College of Education, University of Georgia, she received training in the Educational Research and Measurement. Alexeev's research interests include item response theory, mixture finite models, latent growth models and their applications in mathematics education, cognitive psychology and financial literacy.