



Improving Measurement in the Study of Consumer Finance: Consensus from the NEFE Scale Validation Project

Dee Warmath¹ · Katrina Borowiec² · Ashley B. LeBaron-Black³ · Gary Mottola⁴ · Miranda Reiter⁵ · Kurt Schindler⁶ · Joshua Caraballo⁷

Received: 10 April 2024 / Accepted: 26 November 2024
© The Author(s) 2026

Abstract

Research conducted in the areas of consumer/household finance and financial education is intended to inform policy, programs, and practices that encourage more positive financial decisions and outcomes. The way in which that research is conducted matters. The development, selection, and use of measures is a core element of our research; however, little attention is paid to the topic. The objective of the consensus process reported in this paper was to identify impactful and feasible recommendations for the improvement of measure-related decisions in this field. The work reported here was motivated by a previous study conducted by the National Endowment for Financial Education (NEFE) that evaluated the psychometric properties of existing personal finance measures. The aim of the current study was to identify recommendations for improving the development, selection, and application of measures in the field of consumer/household finance and financial education. A modified Delphi process was used involving a panel of seven experts who used findings from the NEFE study to identify potential recommendations and then engaged in two rounds of discussion and voting. Through this process, a set of 19 recommendations organized into five domains were identified. The intention of this paper is to motivate conversation that can lead to actions in our field that advance our understanding of and ability to impact well-being in the financial domain of life.

Introduction

Consumer finance and education research is intended to inform policy, programs, and practices that encourage consumer behavior and environments that promote greater well-being for individuals and families (Xiao, 2016). The

results of our studies could lead a government or community to invest in a particular program or lead financial advisors, therapists, and counselors to recommend certain behaviors to a client. In some cases, our research informs how an existing program is evaluated, potentially influencing changes in operations or even whether it continues

✉ Dee Warmath
warmath@uga.edu
Katrina Borowiec
katrina.borowiec@bc.edu
Ashley B. LeBaron-Black
lebaronashley@gmail.com
Gary Mottola
gary.mottola@finra.org
Miranda Reiter
mreiter@ttu.edu
Kurt Schindler
kurt.schindler@upr.edu
Joshua Caraballo
jcaraballo@nefe.org

¹ Department of Financial Planning, Housing and Consumer Economics, College of Family and Consumer Sciences, University of Georgia, 205 Dawson Hall, 30602 Athens, Georgia
² Measurement, Evaluation, Statistics, & Assessment (MESA), Lynch School of Education and Human Development, Boston College, Chestnut Hill, USA
³ School of Family Life, College of Family, Home, and Social Sciences, Brigham Young University, Provo, USA
⁴ Office of Investor Education, FINRA, Washington, DC, USA
⁵ School of Financial Planning, College of Human Sciences, Texas Tech University, Lubbock, USA
⁶ Finance Department, University of Puerto Rico at Rio Piedras, San Juan, USA
⁷ National Endowment for Financial Education, Denver, USA

to operate. Therefore, the way in which we conduct our research and the quality of that research matters.

One important decision in the research process has to do with the measures we develop and select for our studies. There is rarely a perfect measure for a given phenomenon, so our choices often, but not always, represent a search for “good enough,” which could include what other researchers are using, or anything available that will suffice. In this study, we are focused on the selection of scales to measure constructs such as money management behavior, planning horizon, financial self-efficacy, and financial well-being as reported to us by the consumer, which can be referred to as subjective measures. Subjective measures can be thought of as those which measure a respondent’s perspective and evaluation of their situation (de Oliveira Cardoso et al., 2023; Sorgente & Lanz, 2019), while objective measures are generally derived from information provided via assessments, ratios, or other quantifiable sources (Tenney & Kalenkoski, 2019). Yes, one could argue that if we used objective measures from administrative records as proxies, when possible, we might need subjective measures less often. However, that argument misses the point that we need to involve the consumer in our research. Some constructs can be best measured or only should be measured subjectively (Jahedi & Mendez, 2014). In fact, there are limitations with using objective measures such as when measuring concepts which are not well-defined (Jahedi & Mendez, 2014). Even when more objective proxies are available, the consumer’s perceptions, attitudes, and reports of behavior matter if our science is going to study the consumer as more than a passive object. Quite often such consumer-focused measures involve scales (i.e., a set of items designed to reflect a given attribute).

Measurement and scale development is a complex field. Thus, it is not possible to provide a comprehensive discussion regarding what constitutes a “good” scale in this paper. Nevertheless, there are several tenets that most measurement experts would highlight. First, the items included in a scale should be consistent with the scale developer’s understanding of the nature of the latent construct being measured (DeVellis, 2017; Jöreskog, 1969). In other words, the scale should be developed in accordance with theory, providing the foundation for content validity and construct validity (Cronbach & Meehl, 1955). Second, a scale should be reliable, meaning that it provides consistent, dependable measures over time (Spearman, 1910). Third, score differences between groups should reflect actual differences in an attribute rather than an underlying bias of the scale (Millsap, 2012).

However, our field, defined generally as the study of consumer/household finance and financial education, suffers from a paucity of valid, reliable, and accepted scales chosen

intentionally for a given study and applied consistently (Grable et al., 2011). When a measure is used repeatedly and, thereby, becomes common practice, it is challenging to put into question the validity and reliability of the measure. This also makes it harder to search for a better way to measure the construct for a given study population or in general. Some constructs in our field were defined by their measure (e.g., financial literacy Warmath, (2022); whereas, theoretically, scales should be developed based on theory about the construct (DeVellis, 2017). When a variety of measures are used for the same construct, often without validation, it is difficult to understand what a single study contributes to the body of knowledge (LeBaron & Kelley, 2021). This hampers our ability to advance knowledge, policy, and theory to improve consumer, household, and community outcomes.

Given the importance of measurement and the lack of attention paid in the field of consumer finance and education research, the National Endowment for Financial Education (NEFE) commissioned a study in 2022 of available measures for selected constructs in the field. Through a literature review, information provided about the validity and reliability of available measures was used to select the best available measure (or measures, where appropriate) for each construct. The focus of their project was on reflective scales assessing social psychological constructs used in the study of consumer/household finance and financial education. The 11 scales selected by NEFE for their project were:

- Brief Money Management Scale (Ksendzova et al., 2017).
- Domain-specific Risk-taking Scale (Blais & Weber, 2006).
- Executive Personal Finance (Spinella et al., 2007).
- Financial Decision-making Ability (Warmath, 2022).
- Financial Self-efficacy (Warmath, 2022).
- Financial Skill (CFPB, 2018).
- Financial Well-being (Netemeyer et al., 2018).
- Material Values Scale (MVS) 15-item (Richins, 2004).
- Personal Savings Orientation (Dholakia et al., 2016).
- Propensity to Overspend on Credit Cards (Sotiropoulos & d’Astous, 2013).
- Propensity to Plan for Money Long-term (Lynch et al., 2010).

In their project, NEFE fielded an online survey containing these scales with a national sample of 993 US adults, oversampling on demographic characteristics such as race, ethnicity, sexual orientation, and gender identity to ensure sufficient data to examine measures overall and by consumer characteristic. Joined by experts in the field, NEFE analyzed the data collected and reviewed the results to determine the best next steps for applying the results to the field.

Although this paper was motivated by the NEFE study, the focus of this paper is not on the results of the exploratory NEFE work. We recognize that there is a much longer list of scales used in our field than the ones selected by NEFE for their study. This paper is also not an attempt to review all scales used in our field. We recognize that there are existing reviews of measures for a particular construct (e.g., financial well-being Bruggen et al., 2017; de Oliveira Cardoso et al. 2023; Sorgente and Lanz 2017), but they primarily strive to consolidate information about a given construct. Instead, the objective of this paper is to identify barriers and opportunities to improve the ways in which measures are developed, selected, and interpreted.

We accomplish this objective through a type of paper rarely (if ever) used in our field: a panel consensus paper. However, consensus papers have been used widely in other established fields such as medicine and health (Fink et al., 1984). The main purpose of consensus papers is to provide solutions to a given problem in a field or profession by finding common ground, oftentimes within a group of experts (Fink et al., 1984). In this paper, we (1) describe our application of this process, (2) share the recommendations that were derived from the process, and (3) suggest ways in which these recommendations could lead to enhanced focus on measurement in our field. Given the paucity of research and conversation on measurement in our field, this consensus-based, modified Delphi approach was warranted. The contribution of our work is a set of recommendations intended to inspire conversation that ideally leads to an increased focus on measurement.

To inform panel consensus, scales from the NEFE study were analyzed in terms of their validity, reliability, and measurement invariance. Three areas of inquiry were considered:

1. Whether the hypothesized relationship between the items and the underlying latent construct fits the new data (confirmatory factor analysis).
2. Whether there is evidence that the items included in the scale measure the same construct (reliability or internal consistency).
3. Whether the measurement properties of a scale are consistent across subgroups of consumers (measurement invariance).

Methods

Participants and Procedures

The procedures used in this study were modeled after an application of the Delphi method to concussion education

(Kroshus et al., 2020). The Delphi method provides an efficient and unbiased approach to identifying collective opinions of experts (Adler & Ziglio, 1996). It is particularly valuable for addressing complex, multidimensional issues where there may be a lack of definitive data or where expert judgment is needed to guide decision-making. The Delphi method is a structured, iterative process that seeks to gather expert opinions through multiple rounds of data collection. The process allows experts to reconsider their views in light of feedback from others, which helps build a consensus over time. A combination of divergent thinking in an exploratory phase and convergent thinking in an evaluative phase serves to first generate, and then refine or prioritize, possible recommendations.

In the context of understanding the intersection between finances and household (family) functioning, the Delphi method is especially appropriate because these areas involve highly interconnected variables, including psychological, social, and financial factors. Current economic conditions—such as rising living costs, inflation, and growing income inequality—place significant stress on families, making it more urgent to understand how financial difficulties affect family dynamics. Moreover, traditional research methods may struggle to capture the full complexity of these relationships, as they often fail to address the emotional, psychological, and behavioral aspects of financial decision-making within families.

The Delphi method fills this gap by involving a panel of seven experts from various fields (e.g., financial education, household finance, family science, and measurement). These experts can offer insights that go beyond basic data analysis, bringing in their knowledge of how finances impact family functioning on multiple levels, including relational dynamics, mental health, and social behavior.

Defining Consensus or Agreement

It is important to place the recommendations made in their intended context. “Consensus” or “agreement” do not mean that the panel is in complete alignment with these recommendations as prescriptions that should be adopted as written and implemented post haste. Instead, we use the terms “consensus” and “agreement” to reflect that we collectively believe that accomplishing some of these recommendations would be beneficial to our science and the groups it serves, in ways that make sense for our field. Our intent is to spark further thought and conversations rather than to make declarations. We also acknowledge that other disciplines (e.g., education, psychology) have done this work for their respective disciplines. Our objective is to increase the attention given to measurement within the discipline of

consumer/household finance and financial education. With that perspective, let us examine how we applied the Delphi approach using phases of exploration and evaluation.

Exploratory Phase

In January of 2024, a total of 31 candidate or suggested recommendations were generated independently by panel members using a Google Sheet. Appendix Table 3 contains these 31 candidate recommendations. These recommendations were discussed at the subsequent panel meeting followed by the first round of voting. During the first round of voting, panel members were asked to rate each recommendation (also referred to as “item” below) on a five-point scale with respect to its impact and feasibility as described below.

Impact: Please rate the likely impact of each of these possible recommendations (if they were to be adopted by the readers). Very impactful, somewhat impactful, not sure, not very impactful, not at all impactful.

Feasibility: Please rate the likely feasibility (or ease) of each of these possible recommendations (if they were to be adopted by the readers). Very feasible/easy to do, somewhat feasible, not sure, not very feasible, not at all feasible/difficult to do.

To provide a simple approach to prioritizing the recommendations for discussion, candidate recommendations received four points if a given rater indicated they were rated as “very impactful” and “very feasible” and an additional one point for every “very impactful” rating regardless of feasibility. The rough logic here was that very impactful and very feasible recommendations should be prioritized given that they represented low-hanging fruit (i.e., impactful ideas that should be very easy to implement. In contrast, items rated as very impactful but not very feasible should be considered due to their potential impact but given lower prioritization due to the perceived challenges of their implementation. The total possible points were 25. The actual distribution of points was 0 to 17 with an average rating of 6.16. Table 1

presents an example of how the points would be applied to a recommendation based on the ratings. Again, these points were simply used to summarize the ratings and prioritize the items for discussion.

The distribution of scores for Round 1 was examined and a cutline was set at a score of four given our objective of producing a meaningful yet parsimonious set of recommendations. With this cutoff, 19 of the 31 items survived Round 1 voting where “survived” means that these items were recommended for Round 2 discussion and voting. The items at or below a score of four were recommended for exclusion. It is important, however, to note that the voting and sorting was used only to guide our attention and not to apply rigid cuts to the recommendations.

Evaluation Phase

At the next panel meeting (February 2024), the 19 items that survived the cutoff point were reviewed first, with edits made to clarify a given recommendation or acknowledge potential barriers to implementation. Then, the 13 items below the cutoff were discussed and panel members were permitted to make a case for inclusion of a dropped recommendation in the second round of voting. Following this discussion, one item was added to the second-round list. This panel meeting also included the initial discussion of suggested recommendation domains, which covered how the recommendations would be categorized or grouped (e.g., offering training in measurement).

The second round of voting used the same rating categories and response options as the first round. In the second round of voting, consensus was achieved with a set of 20 items. However, a recommendation was made that two items be collapsed into a single item given their similarity. This additional change resulted in a final set of 19 recommendations. Panel members provided input into the set of domain names used to organize the recommendations as well as which recommendations would be assigned to each domain. Round 2 comments were used to refine domain names.

Results and Discussion

The recommendations generated through the Delphi process revealed five domains. These domains were initially derived from the Round 1 set of 31 recommendations and then refined during Round 2 voting and discussion. The domains exist to help organize and make sense of the recommendations. A case could easily be made that a given recommendation (or element of that recommendation) should fall

Table 1 Rating example

Rater	Impact	Feasibility	Points earned
A	Very impactful	Very feasible	4
B	Very impactful	Not sure	1
C	Somewhat impactful	Not sure	0
D	Very impactful	Somewhat feasible	1
E	Very impactful	Very feasible	4
Total points			10

Table 2 Consensus recommendations

Domain	Recommendation
Establishing criteria for measurement development, selection, and evaluation	1. Minimum criteria should be established for the process and reporting of measurement quality in studies using established scales. 2. Guidelines should be established for the interpretation and application of measurement invariance findings. 3. Minimum criteria should be established for the process and reporting of measurement quality in studies developing new scales.
Providing incentives for consideration of measurement topics and issues	4. Research funding should be made available to support the development, validation, and continuous improvement of high-quality measures. 5. Journals should encourage publications related to measurement/scale development. 6. Journal editors and associate editors should consider adding measurement topics as an additional paper type that they publish.
Offering training in measurement	7. There should be more training in measurement evaluation and selection offered in graduate school. 8. Measurement invariance methods should be taught in graduate programs. 9. Graduate programs and committees should require graduate students to justify their selection of measures and note the limitations of available measures in their limitations, implications, and recommendations. 10. Researchers should be taught how to evaluate measure quality and interpret their results in light of the quality of measures available. 11. Short courses, pre-conference activities, and other approaches should be implemented to support early-, mid-, and late-career faculty or other researchers in understanding and applying the proper methods to measure selection and application.
Supporting researchers in measure selection	12. Our field should develop resources to support researchers in the best possible measurement selection given what is currently available. 13. Our field should identify and pursue avenues to build understanding and consideration of measurement issues including the measurement selection process and the impact of measure selection on results. 14. Our field should support researchers in developing and using culturally responsive measures (where “culturally responsive” refers to a connection between the participant’s every day lived experience and the measure).
Elevating the role of measurement topics and issues	15. Journal editors should ensure they have editorial review board members with expertise in psychometrics. 16. Journal editors and associate editors should seek reviewers who are qualified to assess measurement issues. 17. As one consideration in the peer review process, journal editors, associate editors, and reviewers should provide feedback on measure quality and the author’s treatment of measure quality. 18. Funders should consider the quality of measurement proposed for program and outcome evaluations. 19. More conversation should be encouraged on the state of measurement in our field.

under one of the other four domains. The five domains (not in priority order) were:

1. Establishing criteria for measurement development, selection, and evaluation.
2. Providing incentives for consideration of measurement topics and issues.
3. Offering training in measurement.
4. Supporting researchers in measure selection.
5. Elevating the role of measurement topics and issues.

Domain 1: Establishing Criteria for Measurement Development, Selection, and Evaluation

This domain suggests that more rigorous methods and criteria be applied to measurement in fields related to consumer financial decision making. There are three recommendations:

Recommendation 1

Minimum criteria should be established for the process and reporting of measurement quality in studies using established scales.

Recommendation 2

Guidelines should be established for the interpretation and application of measurement invariance findings.

Recommendation 3

Minimum criteria should be established for the process and reporting of measurement quality in studies developing new scales.

There was agreement among the panel members that measurement issues are not often given the necessary

consideration in research. Selection of a measure has conceptual and practical considerations (Mokkink et al., 2016). The researcher must understand the construct they seek to measure, identify how others in the field have measured that construct, and decide which measure to use, if more than one measure is available. When a measure is available, especially one used often, it seems that there is little examination of how that measure was developed initially. Researchers might ask themselves several questions including: Were accepted practices for measure development and validation applied? What study population was used? How does that original study population differ from the population examined in their current study? Few studies in our field offer such justification for the measure selected. Examining how a given measure was developed is made difficult in some cases by a lack of information in the scale development paper. Some are missing details on response options used or even in the final item set. Some use a single study with exploratory factor analysis and no confirmatory work. There are few examples in our field in which the measurement properties of a scale developed with one population are examined with another population. We rarely use measures of reliability (e.g., test-retest reliability) beyond Cronbach's alpha (i.e., internal consistency). Measurement invariance testing is very rare in our field. If no measure is available for the construct of interest, the researcher must either develop a measure of their own or find a reasonable proxy. Once a measure is selected, a description of the measure is given, more or less, in the methodology of the paper.

Yet, in many studies, little consideration is given to the ways in which a measure might bias the results of the study. There are cases where only a weak measure exists for what is being studied. Until a better measure is developed, the existing measure may be the only option. However, research conducted with a poor measure may do more harm than good. Perhaps the research should wait for a better measure or at least acknowledge the potential bias introduced in the results. Measurement error impacts effect size, which impacts the power of the study. These dynamics influence the results obtained. If those results are interpreted without attention to these broader dynamics, there is a reasonable chance that our research could be biased. There are relatively few papers in our field that explicitly consider the strengths and weaknesses of the measures they selected when interpreting results. When this discussion does occur, it tends to fall to the limitations section rather than considering how results should be interpreted in light of the measures used.

The recommendations presented here simply call for a shared and transparent sense of what should be done and highlight the need for criteria and guidelines. Developing

guidelines takes time, effort, and input from many different perspectives in our field. Other fields operate with such guidelines and can serve as excellent examples of the process and outcome. The gold standard for measurement in educational and psychological research is the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014). While the *Standards* focus on educational and psychological tests, the basic measurement principles can be adapted to other fields for survey development. Another example can be found in qualitative research where guidelines by study type have been developed and maintained over time (Appllebaum et al., 2018).

Domain 2: Providing Incentives for Consideration of Measurement Topics and Issues

When done well, scale development work is time-consuming and expensive, both in terms of the initial investment in item development and concerning further validation of the items for additional populations and subgroups. Coupled with little funding for measurement-focused projects and the reality that journals in our field rarely publish measurement-focused studies, the lack of focus on measurement is completely understandable. For many researchers, especially an assistant professor on a tenure clock, there are few requirements or incentives to address measurement issues in their research. Even when researchers have the time and resources to address measurement issues, the low bar for measurement quality in the field of consumer/household finance and financial education results in little incentive for them to do so. This domain of three recommendations, thus, seeks to highlight the need for greater incentives from funders, journal editors, academic departments, and other sources.

Recommendation 4

Research funding should be made available to support the development, validation, and continuous improvement of high-quality measures.

Recommendation 5

Journals should encourage publications related to measurement/scale development.

Recommendation 6

Journal editors and associate editors should consider adding measurement topics as an additional paper type that they publish.

Most of the attention paid to measurement is directed at the development of new scales. Relatively little attention is paid to studies that examine the limitations or biases of existing measures or that seek to improve on existing measures. Two exceptions were printed in the *Journal of Consumer Affairs* in 2022 and 2023. One was a commentary on measurement error in financial literacy (Gignac & Ooi, 2022) and the other was a consideration of how treatment of the “don’t know” response in financial literacy measures influences results (Wilmarth et al., 2023). It appears that none of the journals in our field have a measurement-focused article type or special issues inviting papers on measurement-related topics (this *Journal of Family and Economic Issues* being an exception). Conferences or other planned gatherings in our field do not seem to have tracks related to measurement topics.

Domain 3: Offering Training in Measurement

In our field, there seems to be greater emphasis on general statistical methods than on measurement. Yet, the brilliant execution of statistical methods with measures of unrecognized bias can yield inaccurate results. Measurement is an important topic within methods; however, it seems to be a topic rarely covered in our field. The likely result is that researchers are underprepared to select measures and consider the influence of that selection on their results and interpretation. This domain of five recommendations identifies areas in which researcher preparation and continued training might be strengthened.

Recommendation 7

There should be more training in measurement evaluation and selection offered in graduate school.

Recommendation 8

Measurement invariance methods should be taught in graduate programs.

Recommendation 9

Graduate programs and committees should require graduate students to justify their selection of measures and note the

limitations of available measures in their limitations, implications, and recommendations.

Recommendation 10

Researchers should be taught how to evaluate measure quality and interpret their results in light of the quality of measures used.

Recommendation 11

Short courses, pre-conference activities, and other approaches should be implemented to support early-, mid-, and late-career faculty or other researchers in understanding and applying the proper methods to measure selection, application, and development.

These recommendations have two targets: graduate students who are emerging scholars and existing scholars who might need training. The enactment of these recommendations, however, falls to others who can facilitate the training. One challenge with the graduate student recommendations might be the lack of faculty at a given institution who can provide such education and an inability of a thesis or dissertation advisory committee to support such requirements. While the panel considered this challenge, the consensus was that current graduate students represent a potential inflection point. If the statistical methods instructor for a given university does not feel prepared to include such training, a colleague in another department or external expert could provide a guest lecture, or students could take classes in another department on campus or through another university if the institution is part of a consortium. If the advisory committee does not feel prepared for such discussions, perhaps an additional member could be added for this purpose, or an external expert could be consulted. Over time, departments might increasingly recognize the value of measurement expertise within their faculty ranks. For existing researchers, consideration should be given to incentives and barriers to such knowledge acquisition. Our recommendations do not specify how researchers should be taught to measure quality and incorporate measurement issues in interpretation except to offer that conferences might be a useful forum for such conversations. If training options were coupled with (a) new article types; (b) special issues focused on measurement; or (c) new funding opportunities, participation is likely to be high as evidenced by recent conferences introducing new requirements (e.g., the standing room only session when two journals announced new data transparency policies at a recent American Marketing Association conference).

Domain 4: Supporting Researchers in Measure Selection

The general thought underlying this domain of three recommendations is that researchers in our field would benefit from information and tools that support a more rigorous consideration of which measures to use in a given study with a given study population. The current approach to identifying and selecting measures seems to rely on a literature search to see what has been used or consultation of a handbook that lists existing scales. One challenge in the current approach is that important information in guiding measure selection is often missing from those sources. As a result, researchers who might want to make measurement decisions more intentionally find it difficult to obtain the information and options necessary to support such decisions. While other domains address the need for increased motivation on the part of the researchers to search for the “right” measure for their study and support structures to guide them, this domain addresses the need to make that search easier.

Recommendation 12

Our field should develop resources to support researchers in the best possible measurement selection given what is currently available.

Recommendation 13

Our field should identify and pursue avenues to build understanding and consideration of measurement issues and the importance of measurement selection.

Recommendation 14

Our field should support researchers in developing and using culturally responsive measures (where “culturally responsive” refers to a connection between the participant’s everyday lived experience and the measure).

The field of consumer/household finance and financial education is not the only field struggling with measurement; we are just one field that to date has not addressed the issue in ways that support researcher decisions. The health fields are an example of how we might support the desired behaviors when it comes to measurement selection and application. In health research, the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) exists to improve measurement in the study of health-related topics (Mokkink et al., 2016). The

team leading the initiative has expertise in epidemiology, psychometrics, qualitative research, and health care, with an emphasis on measurement. The goals of this initiative are to (a) advance the science and application of health outcome measurement; (b) develop new and update existing methodology and practical tools for the selection and use of outcome measurement instruments for research and clinical practice; (c) monitor and maintain the scientific quality of COSMIN tools; (d) encourage widespread adoption of the COSMIN methodology; and (e) call for standardization of outcomes and outcome measurement instruments by developing Core Outcome Sets (COS) and COS methodology. COSMIN exists to define what is required for a quality measure and a quality methodology to develop such measures. They also provide advice on measure selection, including an interactive website to support such decisions (<https://www.cosmin.nl/>). How might our field identify ways to validate and standardize the measures we use while acknowledging the different cultural experiences of our study populations?

Domain 5: Elevating the Role of Measurement Topics and Issues

The final domain of recommendations has to do with increasing the importance and discussion of measurement-related topics. While this conversation needs to take place in multiple forums involving all of us, we believe that there are a few opportunities to support the initiation of those conversations. For that reason, our five recommendations are directed toward the editorial teams of our journals and the funders in our field, with a final recommendation addressing the need for conversation in general.

Recommendation 15

Journal editors should ensure they have editorial review board members with expertise in psychometrics.

Recommendation 16

Journal editors and associate editors should seek reviewers who are qualified to assess measurement issues.

Recommendation 17

As one consideration in the peer review process, journal editors, associate editors, and reviewers should provide feedback on measure quality and the author’s treatment of measure quality.

Recommendation 18

Funders should consider the quality of measurement proposed for program and outcome evaluations.

Recommendation 19

More conversation should be encouraged on the state of measurement in our field.

The primary forum for conversations in an academic field is the collection of journals that publish our work. Journals and their editors have a responsibility to advance the field. Policies regarding data transparency and study replicability are two recent examples of where an editor might act for this purpose. Recent efforts from the *Journal of Marketing Research* and the *Journal of Marketing* illustrate how journals and editors are getting involved in defining expectations for researchers in their field. Through various channels, the editors of these journals are communicating expectations for data and code sharing to increase transparency and how the desired (soon to be required) researcher behaviors will be supported by these journals. One of several elements of an editor's responsibility is protecting the integrity of the research. If our research is intended to inform public policy, consumer behavior, financial education, or another well-being-enhancing effort, research integrity is mandatory.

However, journals are not the only forum. Such conversations could occur in our organizations, departments, and other settings where researchers gather. How might leaders in these settings encourage conversations related to measurement? Looking again at the health field, we see an example in an organization known as ISOQOL (<https://www.isoqol.org>) with a stated mission to "advance health-related quality of life research." One area of emphasis in this organization is measurement and the concept of measuring what matters. Our field might find inspiration in this organization's acknowledgement of the importance of measurement in advancing knowledge in their field.

The NEFE panel agreed that because we see these recommendations as the start of an important conversation, no requirements should be established related to measurement at this point in the process. Our intention is to encourage researchers, journal editors, funders, organizations, and others to begin the conversation. Once a study has been conducted, there is little that can be done to change the measures. There is, however, an opportunity to provide feedback on the use of those measures. Encouraging reviewers to provide feedback on the measures selected and suggestions

for how they might recognize bias in their interpretation of results can help begin that conversation at the primary forums for our field.

Conclusion

The results of this application of the Delphi method to the state of measurement in the field of consumer finance suggest that measurement has not been a major focus in our field. The fact that 19 recommendations could be identified through two rounds of exploration and evaluation indicates that there are many important topics related to measurement that have not been considered comprehensively. Our recommendations include (a) elevating the importance of measurement; (b) preparing researchers for the development and selection of valid, reliable, and culturally responsive measures, and (c) providing resources and incentives to encourage more responsible behaviors in study design and interpretation. Remember that these recommendations are not prescriptions for action but suggestions of topics that should be considered. The recommendations speak to a wide range of people and positions in our field including editors, associate editors, reviewers, funders, graduate students, graduate programs, conference organizers, and anyone who conducts research in the areas of consumer finance and financial education, defined broadly.

While it is important for our field to use consumer-facing measures, it is equally important that we develop, select, and apply those measures thoughtfully and carefully. In this report, we do not attempt to solve the issues with the measures themselves, but to begin to discuss the gaps in our field that might prevent greater effort related to measurement. Many more people need to get involved in those conversations and potential actions than the seven people convened on the panel by NEFE. The recommendations we make here are likely to need refinement. There are also likely to be other recommendations that we did not consider. Our hope is that this work initiates conversations that lead to the identification of impactful and feasible actions that advance our field through improvements in measurement.

Appendix

See Table 3

Table 3 Round 1 candidate items and scores

Candidate Item	Rating
Researchers should be taught how to evaluate measure quality and interpret their results in light of the quality of measures available.	17
Research funding should be made available to support the development and validation of high-quality measures.	12
Editors and associate editors should consider adding measurement topics as an additional paper type that they publish.	11
Guidelines should be established for the interpretation of measure invariance findings and when a particular scale should (not) be used with a given population.	11
Journal editors should ensure they have editorial review board members with expertise in psychometrics.	11
Journals should encourage publications related measurement/scales development.	11
Minimum criteria should be established for the process and reporting of measurement quality in studies using established scales.	11
Graduate programs and committees should require graduate students to justify their selection of measures and note the limitations of available measures in their limitations, implications, and recommendations.	9
Measurement invariance methods should be taught in graduate programs.	8
Editors and associate editors should seek reviewers who are qualified to assess measurement issues.	7
Minimum criteria should be established for the process and reporting of measurement quality in studies developing new scales.	7
Our field should identify and pursue avenues to build awareness of measurement issues and the importance of measurement selection.	7
Research funding should be made available to support continuous improvement of existing measures.	7
Researchers need guidance to support the best possible measurement selection given what is currently available.	7
There should be more training in measurement evaluation and selection offered in graduate school.	7
Whether a measure is culturally responsive should be a primary factor in a researcher's determination of whether to use that measure in their study.	7
Funders should consider the quality of measurement proposed for program and outcomes evaluations.	6
Short courses, pre-conference activities, and other approaches should be implemented to support early, mid and late-career faculty in understanding and applying the proper methods to measure selection and application.	6
More conversation is needed on the state of measurement for our constructs.	6
Editors, associate editors, and reviewers should include measure quality and the author's treatment of measure quality as one consideration in determining whether to accept a paper.	4
Researchers need a way to identify the best available scale for their constructs, research setting, and populations of interest.	4
There should be more training in measurement development methods offered in graduate school.	4
Early career and older should have access to training in measurement evaluation and selection.	3
Our field should prioritize the development and validation of better measures.	3
Consideration of measurement invariance for the subgroups in a given study should be required.	1
Early career and older should have access to training in measurement development methods.	1
Our field needs a way to identify and communicate poor measures that should no longer be used.	1
Our field should seek approaches to having a broader conversation of measurement-related topics.	1
We should acknowledge the challenges with the measures we have available (e.g., overlapping concepts).	1
Assessment of measurement invariance for a minimum set of subgroups should be required in scale development.	0
Minimum criteria should be established for the subgroups on which measurement invariance is established for a given study.	0

Acknowledgements Any opinions provided herein are those of the authors and do not necessarily reflect the views of the institutions with which the authors are affiliated.

Funding This research was funded by a grant from the National Endowment for Financial Education.

Data Availability There is no data associated with this consensus paper.

Declarations

Conflict of interest The authors do not have any conflicts of interest to disclose that are directly or indirectly related to the work submitted for publication.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the

Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

References

- Adler, M., & Ziglio, E. (1996). *Gazing into the oracle: The Delphi method and its application to social policy and public health*. Jessica Kingsley.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*.

- American Educational Research Association. https://www.teststandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf
- Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, 73(1), 3. <https://doi.org/10.1037/amp0000389>
- Blais, A. R., & Weber, E. U. (2006). A domain-specific risk-taking (DOSPERT) scale for adult populations. *Judgment and Decision making*, 1(1), 33–47. <https://doi.org/10.1017/S1930297500000334>
- Consumer Financial Protection Bureau (CFPB) (2018). Measuring Financial Skill. <https://www.consumerfinance.gov/data-research/research-reports/measuring-financial-skill/>
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281–302. <https://doi.org/10.1037/h0040957>
- de Oliveira Cardoso, N., Markus, J., de Lara Machado, W., & Guilherme, A. A. (2023). Measuring financial well-being: A systematic review of psychometric instruments. *Journal of Happiness Studies*, 24(8), 2913–2939.
- DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). Sage.
- Dholakia, U., Tam, L., Yoon, S., & Wong, N. (2016). The ant and the grasshopper: Understanding personal saving orientation of consumers. *Journal of Consumer Research*, 43(1), 134–155. <https://doi.org/10.1093/jcr/ucw004>
- Fink, A., Kosecoff, J., Chassin, M., & Brook, R. H. (1984). Consensus methods: Characteristics and guidelines for use. *American Journal of Public Health*, 74(9), 979–983. <https://doi.org/10.2105/AJPH.74.9.979>
- Gignac, G. E., & Ooi, E. (2022). Measurement error in research on financial literacy: How much error is there and how does it influence effect size estimates? *Journal of Consumer Affairs*, 56(2), 938–956. <https://doi.org/10.1111/joca.12417>
- Grable, J. E., Archuleta, K. L., & Nazarinia, R. R. (2011). Financial planning and counseling scales. *Springer*. <https://doi.org/10.1007/978-1-4419-6908-8>
- Jahedi, S., & Méndez, F. (2014). On the advantages and disadvantages of subjective measures. *Journal of Economic Behavior & Organization*, 98, 97–114. <https://doi.org/10.1016/j.jebo.2013.12.016>
- Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34(2), 183–202.
- Kroshus, E., Cameron, K. L., Coatsworth, J. D., D'Lauro, C., Kim, E., Lee, K., & Hainline, B. (2020). Improving concussion education: Consensus from the NCAA-department of defense mind matters research & education grand challenge. *British Journal of Sports Medicine*, 54(22), 1314–1320. <https://doi.org/10.55460/6UWZ-D7UU>
- Ksendzova, M., Donnelly, G. E., & Howell, R. T. (2017). A brief money management scale and its associations with personality, financial health, and hypothetical debt repayment. *Journal of Financial Counseling and Planning*, 28(1), 62–75. <https://doi.org/10.1891/1052-3073.28.1.62>
- LeBaron, A. B., & Kelley, H. H. (2021). Financial socialization: A decade in review. *Journal of Family and Economic Issues*, 42, 195–206. <https://doi.org/10.1007/s10834-020-09736-2>
- Lynch Jr, J. G., Netemeyer, R. G., Spiller, S. A., & Zammit, A. (2010). A generalizable scale of propensity to plan: The long and the short of planning for time and for money. *Journal of Consumer Research*, 37(1), 108–128. <https://doi.org/10.1086/649907>
- Millsap, R. E. (2012). *Statistical Approaches to Measurement Invariance*. Routledge.
- Mokkink, L. B., Prinsen, C. A., Bouter, L. M., de Vet, H. C., & Terwee, C. B. (2016). The COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) and how to select an outcome measurement instrument. *Brazilian Journal of Physical Therapy*, 20, 105–113. <https://doi.org/10.1590/bjpt-rbf.2014.0143>
- Netemeyer, R. G., Warmath, D., Fernandes, D., & Lynch Jr, J. G. (2018). How am I doing? Perceived financial well-being, its potential antecedents, and its relation to overall well-being. *Journal of Consumer Research*, 45(1), 68–89. <https://doi.org/10.1093/jcr/ucx109>
- Richins, M. L. (2004). The material values scale: Measurement properties and development of a short form. *Journal of Consumer Research*, 31(1), 209–219. <https://doi.org/10.1086/383436>
- Sorgente, A., & Lanz, M. (2017). Emerging adults' financial well-being: A scoping review. *Adolescent Research Review*, 2, 255–292. <https://doi.org/10.1007/s40894-016-0052-x>
- Sorgente, A., & Lanz, M. (2019). The multidimensional subjective financial well-being scale for emerging adults: Development and validation studies. *International Journal of Behavioral Development*, 43(5), 466–478. <https://doi.org/10.1177/0165025419851859>
- Sotiropoulos, V., & d'Astous, A. (2013). Attitudinal, self-efficacy, and social norms determinants of young consumers' propensity to overspend on credit cards. *Journal of Consumer Policy*, 36, 179–196. <https://doi.org/10.1007/s10603-013-9223-3>
- Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, 3, 171–195.
- Spinella, M., Yang, B., & Lester, D. (2007). Development of the executive personal finance scale. *International Journal of Neuroscience*, 117(3), 301–313. <https://doi.org/10.1080/00207450500534043>
- Tenney, J. A., & Kalenkoski, C. M. (2019). Financial ratios and financial satisfaction: Exploring associations between objective and subjective measures of financial well-being among older Americans. *Journal of Financial Counseling and Planning*, 30(2), 231–343. <https://doi.org/10.1891/1052-3073.30.2.231>
- Warmath, D. (2022). Financial decision making. Report to the Australian Securities and Investment Commission.
- Wilmarth, M. J., Kim, K. T., & Pak, T. Y. (2023). What do we really know about don't know? Re-assessing the measurement of financial knowledge. *Journal of Consumer Affairs*, 57(4), 1623–1649. <https://doi.org/10.1111/joca.12563>
- Xiao, J. J. (2016). *Handbook of consumer finance research*. Springer.