Profil 10:

Herausforderungen und Gestaltungsfragen für die berufliche Bildung

Digitale Festschrift
für SUSAN SEEBER

Elisa WAGNER[1], Eveline WUTTKE[2] & Roland HAPP[1]
([1]Universität Leipzig, [2]Universität Frankfurt)

Valide Messung von Financial Literacy – Überarbeitung und Analyse eines Situational Judgement Tests

Online unter:

https://www.bwpat.de/profil10_seeber/wagner_etal_profil10.pdf

in

***bwp@* Profil 10** | November 2024

Herausforderungen und Gestaltungsfragen für die berufliche Bildung

Hrsg. v. **Christian Michaelis, Robin Busse, Eveline Wuttke & Bärbel Fürstenau**

www.bwpat.de | ISSN 1618-8543 | *bwp@* 2001–2024

bwp@

www.bwpat.de

Herausgeber von *bwp@*: Karin Büchter, Franz Gramlinger, H.-Hugo Kremer, Nicole Naeve-Stoß, Karl Wilbers & Lars Windelband

Berufs- und Wirtschaftspädagogik - *online*

ELISA WAGNER[1], EVELINE WUTTKE[2] & ROLAND HAPP[1]
([1]Universität Leipzig, [2]Universität Frankfurt)

Valide Messung von Financial Literacy – Überarbeitung und Analyse eines Situational Judgement Tests

Abstract

Für die Erfassung von Financial Literacy (FL) gibt es bereits eine Vielzahl standardisierter Messinstrumente, die sich allerdings insbesondere der Limitation stellen (müssen), dass sie nicht über die Erfassung von Wissen und darauf operierendem Verstehen hinausgehen. Das reicht nicht aus, um valide Aussagen über finanzbezogenes Handeln zu treffen. Der Beitrag stellt die Konstruktion und im Speziellen die Überarbeitung eines Situational Judgement Tests (SJT) zur Erfassung von FL junger Erwachsener vor. SJT erheben den Anspruch, möglichst realitätsnah das Verhalten von Proband:innen zu messen. Allerdings ist die Konstruktion der Aufgaben insbesondere mit Blick auf die psychometrische Güte der Messinstrumente herausfordernd. Der Beitrag geht auf empirische Befunde zu 885 jungen Erwachsenen ein. Neben der Faktorstruktur und Reliabilität wird ein Augenmerk auf die Frage geworfen, ob die Testwerte valide Testwertinterpretationen zulassen. Als Außenkriterium dient das Kaufverhalten der Proband:innen. Die Befunde unterstreichen die Herausforderungen, die bei der Konstruktion und Auswertung von SJT bestehen. Es sind weitere Forschungsansätze notwendig, um künftig auf valide Messinstrumente zurückgreifen zu können.

Valid Measurement of Financial Literacy – Revision and Analysis of a Situational Judgment Test

There are already a large number of standardized measurement instruments for assessing financial literacy (FL), but these (have to) face the limitation that they do not go beyond the assessment of knowledge and understanding based on this. This is not enough to make valid statements about finance-related actions. This article presents the construction and, in particular, the revision of a situational judgement test (SJT) to assess the FL of young adults. SJT claim to assess the behavior of test subjects as realistically as possible. However, the construction of the tasks is challenging, particularly with regard to the psychometric quality of the measurement instruments. The article addresses empirical findings on 885 young adults. Attention is paid to the factor structure and reliability as well as to the question of whether the test scores allow valid test score interpretations. The purchasing behavior of the test subjects serves as an external criterion. The findings emphasize the major challenges that exist in the construction and evaluation of SJT. Further research is necessary in order to have valid measurement instruments available for this purpose.

***Schlüsselwörter**: Financial Literacy, Situational Judgement Test, Validierung, messmethodische Qualität*

***Keywords**: Financial Literacy, Situational Judgement Test, validation, methodological quality*

1 Die Notwendigkeit des validen Messens von Financial Literacy

Die valide Messung von Kompetenzen ist seit vielen Jahren ein Anliegen von Susan Seeber (z. B. Seeber/Nickolaus 2010, Seeber et al. 2014). Ein zentraler Fokus ist ihre Forderung – speziell bezogen auf berufliche Bildung – nicht nur (träges) Wissen abzufragen, sondern komplexe und authentische Tests zu entwickeln, die es erlauben, Rückschlüsse auf die Kompetenzen der jeweiligen Zielgruppe zu ziehen (z. B. Wuttke et al. 2022). Wir knüpfen mit diesem Beitrag an ihre Forschung an und nehmen dabei die Financial Literacy (FL) bei jungen Erwachsenen in den Blick.

In den vergangenen Jahren hat eine Vielzahl von Studien gezeigt, dass das Wissen und Können junger Erwachsener im Bereich der FL mangelhaft ist (z. B. Happ/Förster 2019; Lusardi/Mitchell/Curto 2010). Das betrifft auch solche Länder, die über gut entwickelte Wirtschafts- und Finanzmärkte verfügen und stellt auch in Deutschland eine Herausforderung dar, der sich in erster Linie nicht jedes Individuum selbst, sondern insbesondere das Bildungswesen stellen muss (vgl. BMF/BMBF 2023; Bucher-Koenen/Knebel 2021; Lusardi 2019). Um adressatengerechte Bildungsmaßnahmen zu entwickeln, muss zunächst zuverlässig ermittelt werden, über welche Kompetenzen junge Erwachsene verfügen, um mit entsprechenden Bildungsmaßnahmen anknüpfen zu können.

Zur Messung der FL gibt es bereits eine Vielzahl von teils sehr unterschiedlichen Ansätzen und Instrumenten (vgl. Förster/Happ/Molerov 2017; Klapper/Lusardi/Van Oudheusden 2015; Lusardi/Mitchell 2008 bzw. Lusardi/Mitchell 2011; OECD 2020; Walstad/Rebeck 2017; Walstad/Rebeck 2018; Wuttke/Aprea 2018 bzw. Wuttke/Siegfried/Aprea 2020 etc.), die verschiedene Aspekte (Wissen, Fähigkeiten, Fertigkeiten, Motivation) zu unterschiedlichen Inhaltsbereichen (Budgetierung, Versicherungen, Anlageentscheidungen, Zinsrechnung etc.) messen. Die in den Studien eingesetzten Messinstrumente sind aus mehreren Gesichtspunkten umstritten. Manche Aufgaben (insb. die sog. ‚Big Three' von Lusardi (2008)) sind sowohl aus inhaltlicher als auch messtheoretischer Sicht problematisch, da sie nur eine sehr kleine Schnittmenge der theoretischen Modellierung von FL abdecken.[1]

Die Aufgaben, die vom US-Amerikanischen Council for Economic Education (CEE) für verschiedene Altersstufen konstruiert wurden, werden in der Praxis im Original und in länderspezifischen Adaptionen häufig eingesetzt (für Deutschland und Korea: Happ et al. 2022; für die Niederlande Amagir/Wilschut/Groot 2018; für die USA Walstad/Rebeck 2017). Die Instrumente des CEE decken mit 40 bis 45 Aufgaben je Messinstrument viele Inhaltsbereiche ab, aber mit Blick auf die Breite des Konstrukts bei der OECD (2020) gehen diese kaum über das Erfassen von Wissen und darauf operierendem Verstehen hinaus. Sie können daher als wissensbasiert charakterisiert werden.

[1] Es gibt eine Vielzahl von FL Definitionen. Atkinson und Messy (2012, 14) definieren FL als „combination of awareness, knowledge, skill, attitude and behaviour necessary to make sound financial decisions and ultimately achieve individual financial wellbeing." Diese Publikation ist auch noch für die heutige Konzeptualisierung von FL durch die OECD (2020) prägend, an der wir uns in diesem Beitrag orientieren.

Dass es eine Diskrepanz zwischen Wissen und Handeln eines Individuums gibt, ist bekannt (z. B. Mandl/Gerstenmeier 2000). Träges Wissen wird als Wissen definiert, welches Menschen zwar abrufen und wiedergeben, nicht aber in relevanten Anwendungssituationen in Handlungen umsetzen können (vgl. Gruber/Renkl 2000; Renkl 1996; Whitehead 1929). Viele (Wissens-)Tests, die für ganz verschiedene Inhaltsbereiche eingesetzt werden, können deshalb nur bedingt das tatsächliche Handeln von Menschen vorhersagen. Nicht zuletzt deshalb finden sich nur wenige Studien, die überhaupt Auskunft über die (prognostische) Vorhersagekraft eines FL Tests geben (z. B. in Schmeiser/Seligman 2013; Rieger 2020). Abgesehen vom Phänomen des trägen Wissens spielen weitere Variablen bei der Umsetzung von Wissen in Handeln eine Rolle, die dazu führen, dass ungünstige Entscheidungen getroffen werden, welche nicht rationaler Natur sind. Das Arrested Deployment Model von Carmel, Leiser und Spivak (2020) verdeutlicht dies, indem es den Einfluss weiterer Faktoren (psychologische Aspekte, Aufgabenmerkmale, äußere Einwirkungen) auf solche Prozesse darstellt.

Ein von Wuttke und Aprea (2018) entwickelter Situational Judgement Test (SJT) setzt an der beschriebenen Herausforderung an. Er misst FL verhaltensnah und wurde nach einem ersten Einsatz 2016 und 2017 auf Basis mehrerer Studien überarbeitet, um aus messmethodischer Sicht Verbesserungen zu erzielen. Die Validität als Gütekriterium für die psychometrische Qualität der Daten spielt eine zentrale Rolle. Auf Basis der Standards for Educational and Psychological Testing der AERA, APA und NCME (2014) ist eine valide Interpretation der Testwerte nur möglich, wenn belastbare Evidenz bei einer Vielzahl von Kriterien (bezogen auf den Testinhalt, den Antwortprozessen, der internen Struktur, der Beziehung zu anderen Variablen oder den Konsequenzen der Testung, s. ausführlicher in Kapitel 3.3.4) vorliegt.

Der Beitrag hat das Ziel, die Überarbeitung des Testinstruments vorzustellen und einen Einblick in den Validierungsprozess zu geben. Folgende Forschungsfragen werden beantwortet:

1) Wie wurde der SJT konstruiert und welche Veränderungen wurden nach einer ersten Testung am SJT vorgenommen?
2) Welche (interpretierbare) Faktorstruktur bilden die Testitems nach der Überarbeitung des Instrumentes?
3) Wie sind die Itemschwierigkeiten und Itemtrennschärfen für den SJT zu bewerten?
4) Messen die Skalen das zugrundeliegende Konstrukt reliabel?
5) Welcher Zusammenhang besteht zwischen den erreichten Punkten im SJT und einem Außenkriterium als ein Kriterium für Aussagen zur validen Testwertinterpretation?

Zur Beantwortung der Fragen stellen wir in Kapitel 2 die Grundprinzipien zu dem Instrument vor. Wir gehen darauf ein, wieso diese als Alternative zu Wissensabfragen (klassische Multiple- oder Single-Choice Aufgaben) angesehen werden können. Im daran anschließenden Kapitel 3 werden die Anlässe aufgezeigt, die zu einer Überarbeitung des Instruments geführt haben. Kapitel 4 stellt die empirische Studie vor, die auf Basis der überarbeiteten Version in 2023 und 2024 durchgeführt wurde. Hierbei werden die Gütekriterien der Testung aus 2016/2017 den von 2023/2024 gegenübergestellt und messmethodisch bewertet. Der Beitrag endet mit einer kritischen Würdigung des Vorgehens und einem Ausblick.

2 Situational Judgement Tests als Alternative zu Wissenstests

Es besteht weitestgehend Einigkeit darüber, dass das deutsche Bildungssystem die finanzielle Allgemeinbildung fördern sollte (vgl. BMF/BMBF 2023). Dies impliziert jedoch auch, dass geeignete Instrumente zur Beurteilung des Förderungsbedarfs und später des Lernerfolgs vorhanden sind oder entwickelt werden. Ziel dabei sollte sein, nicht nur vereinzelte Wissensteile zu erfassen, sondern ganzheitlich Kompetenzen von Lernenden in den Blick zu nehmen. Bei der Messung muss zudem die psychometrische Qualität der jeweiligen Instrumentarien sichergestellt sein. Viele der bisher verfügbaren Instrumente werden jedoch diesen Anforderungen – wenn überhaupt – nur in Teilen gerecht. Häufig beruhen sie auf Selbsteinschätzungen und vernachlässigen dabei, dass Menschen dazu neigen, ihre Kompetenzen zu überschätzen (vgl. Nickolaus 2010, 485). Tests, die bislang in der Forschung verwendet werden, haben zudem oft Grenzen: (1) sie testen überwiegend Wissen, statt Kompetenzen, (2) sie sind nicht immer auf authentische Probleme der jeweiligen Zielgruppe bezogen und (3) sie vernachlässigen nicht-kognitive Aspekte, wie z. B. die Fähigkeit zu Belohnungsaufschub oder ethisch-moralische Überlegungen im Zusammenhang mit finanziellen Entscheidungen (z. B. Aprea 2012; Breuer 2016; Huston 2010; Remund 2010).

Werden solche Testergebnisse herangezogen, um Verhalten in späteren realen Situationen vorherzusagen, wird häufig eine Lücke zwischen Wissen und Handeln festgestellt (vgl. Gruber/Renkl 2000; Renkl 1996; Whitehead 1929). Testergebnisse, die auf Wissensabfragen beruhen und sich zudem auf wenige Inhalte beziehen, sagen nicht notwendigerweise späteres Verhalten vorher (vgl. Schmeiser/Seligman 2013). SJT haben dagegen den Anspruch, Verhalten in Anforderungssituationen gut vorherzusagen. Sie erfassen Aspekte, die von herkömmlichen Testformen nicht abgebildet werden (vgl. Kahmann 2014, 49). Bei der Testung mit SJT werden den Proband:innen realitätsnahe, hypothetische Situationen oder Szenarien präsentiert und sie werden aufgefordert, die angemessenste Antwort zu identifizieren oder die Antworten in der Reihenfolge zu ordnen, die sie für am besten geeignet halten (z. B. Kahmann 2014; McDaniel/Nguyen 2001; Whetzel/McDaniel 2009). SJT bestehen deshalb klassischerweise aus Situationen (Itemstämme) und Handlungsoptionen (Itemantworten bzw. kurz ‚Items') (vgl. McDaniel/Nguyen 2001, 104). Es wird angenommen, dass SJT das prozedurale kontextspezifische Wissen und die situative Entscheidungsfähigkeit der Teilnehmer messen (vgl. Kahmann 2014, 49). Dieser Ansatz scheint auch für die Messung von FL geeignet, da er realitätsnahe Probleme und Situationen des täglichen Lebens verwendet, um kompetentes Verhalten in ähnlichen Situationen zu messen, in denen finanzielle Entscheidungen getroffen werden.

Der von Wuttke und Aprea (2018) entwickelte SJT wurde für die Zielgruppe der jungen Erwachsenen konzipiert, da diese Altersgruppe erstmals selbstbestimmt Finanzentscheidungen treffen muss, welche unter Umständen mit weitreichenden Konsequenzen verbunden sind (z. B. die Entscheidung, frühzeitig eine Berufsunfähigkeitsversicherung abzuschließen). Der Test sollte deshalb Situationen und Handlungsoptionen enthalten, die nah an der Lebensrealität dieser Zielgruppe sind.

Dem Test liegt ein zweidimensionales Kompetenzmodell zugrunde. (1) Die inhaltliche Dimension lässt sich in individuelle (Entscheidungen im persönlichen wirtschaftlichen Umfeld) und systembezogene Ausprägungen (gesamtwirtschaftliche Anforderungen) von FL einteilen. (2) Die zweite Dimension unterscheidet Kompetenzfacetten, nämlich kognitive (Wissen, Fähigkeiten, Fertigkeiten) und non-kognitive persönliche Ressourcen (emotionale, motivationale, volitionale Aspekte sowie soziale Werte/Normen). In der Kombination ergeben sich vier Kompetenzbereiche: individuell-kognitiv, individuell-non kognitiv, systemisch-kognitiv und systemisch-non kognitiv (vgl. Aprea et al. 2015, 13f.; Leumann et al. 2016, 23f.). Der entwickelte SJT legt den Fokus auf das Planen und Verwalten alltäglicher Geldangelegenheiten (s. Tabelle 1, Facette 2)[2] – also einer von acht FL Facetten, die dem individuell-kognitiven Kompetenzbereich zuzuordnen ist (vgl. Wuttke/Aprea 2018, 275). Die weiteren Kompetenzfacetten sind in Tabelle 1 ersichtlich.

Tabelle 1: (Sub-)Facetten des individuell-kognitiven Kompetenzbereichs

Facetten		Subfacetten	
1	Geld verdienen		
2	Alltägliche Geldangelegenheiten planen und verwalten	a	Die eigenen Einnahmen einschätzen
		b	Ausgaben in Abstimmung mit den eigenen Bedürfnissen/Möglichkeiten planen
		c	Budget aufstellen und überprüfen
		d	Geld ausgeben
		e	Kurzfristige Geldreserven anlegen
		f	Bank- und Finanzdienstleistungen des täglichen Bedarfs nutzen
3	Geld sparen/Vermögen bilden		
4	Geld leihen/Kredit aufnehmen		
5	Altersvorsorge treffen		
6	Versicherungen abschließen		
7	Überschuldung vermeiden		
8	Informations- und Beratungsangebote zu Geld- und Finanzfragen nutzen		

Anmerkung. Quelle: Wuttke und Aprea (2018, 275).

In der ersten Version enthielt der Test für die Facette des Planens und Verwaltens alltäglicher Geldangelegenheiten 11 Situationen mit insgesamt 46 Items, die sich wie folgt auf drei Subfacetten verteilen:

a Bewertung der eigenen Einnahmen (4 Situationen, je 4 Items);

[2] Grund für die Wahl dieser Facette ist, dass sie eine zentrale Facette ist, die auf jeden Fall beherrscht werden muss. Auch wenn die jungen Erwachsenen vielleicht noch nicht mit gesellschaftlichen Anforderungen (Bewertung von Wirtschaftspolitik, Wahlverhalten) konfrontiert sind und vielleicht auch noch kein zu verwaltendes Einkommen haben, müssen sie alltägliche Geldangelegenheiten sinnvoll verwalten können (Geldzuwendungen der Eltern, Entscheidungen, wie sie die knappen Mittel verwalten etc.). Die Schwerpunktsetzung auf die Facette 2 kann zudem damit begründet werden, dass über die Analysen weniger die Zusammenhänge zwischen den 8 Facetten im Vordergrund stehen. Stattdessen steht die Konstruktion eines SJT für eine Facette im Fokus, was – wie der Beitrag zeigt – bereits eine große Herausforderung darstellt.

b Planung der Ausgaben in Übereinstimmung mit den eigenen Bedürfnissen und Möglichkeiten/Einnahmen (3 Situationen: Situationen 5 und 6 mit 4 Items, Situation 7 mit 6 Items) und

c Aufstellen eines Budgets (4 Situationen, je 4 Items).

Die Frage nach der Fähigkeit zum Belohnungsaufschub war in einige der Items eingebettet, wenn es z. B. darum ging, ggf. ein neues Smartphone zu kaufen, ohne aktuell genügend Mittel zur Verfügung zu haben. Die Testteilnehmenden werden für jede Situation gefragt, wie wahrscheinlich es ist, dass sie sich so verhalten würden, wie dies in den einzelnen Handlungsoptionen beschrieben wird. Dies schätzen sie auf einer fünfstufigen Likert-Skala ein, die von ‚sehr wahrscheinlich' bis ‚sehr unwahrscheinlich' reicht. Likert-Skalen haben den Vorteil, dass jede Handlungsoption einzeln eingeschätzt, die Gefahr ipsativer Daten dadurch minimiert und insgesamt eine bessere Reliabilität erreicht wird (vgl. Weekley/Ployhart/Holtz 2014, 176).

In der Regel geht es bei SJT Aufgaben nicht darum, möglichst viele richtige Antworten zu erzielen. Vielmehr soll geprüft werden, ob jemand eine effektive Reaktion in Bezug auf eine dargestellte Situation zeigt (vgl. Muck 2013, 186). Welche Reaktion als ‚effektiv' oder eher ‚ineffektiv' bewertet wird, richtet sich im betrachteten Fall nach der Meinung verschiedener Expert:innen aus dem FL Bereich, die auch hinzugezogen wurden, als die Handlungsoptionen im Rahmen der Testentwicklung validiert wurden. Ferner musste über das Scoring entschieden werden. Im vorliegenden Fall wird dies in Anlehnung an das Vorgehen von Mumford et al. (2008) durchgeführt, d. h. es werden abgestuft Punkte für jede Antwort verteilt und dann die Summe aller Punktzahlen für die effektiven Antworten und aller umcodierten Punktzahlen für die ineffektiven Antworten gebildet. Bei der Evaluation des Tests wurde bei einigen Situationen und Items ein Überarbeitungsbedarf festgestellt, der im Folgenden in den Blick genommen wird.

3 Methode

3.1 Messmethodische Befunde aus dem ersten Testeinsatz (2016/2017) und Überarbeitung des Instruments

Nach der ersten Testanwendung im Jahr 2016/2017 (N = 206) wurden Faktoren-, Item- und Reliabilitätsanalysen durchgeführt (vgl. Wuttke/Aprea 2018). Die Ergebnisse zeigen eine Drei-Faktoren-Lösung mit guter Interpretierbarkeit (Faktor 1: Überblick/Kontrolle der eigenen finanziellen Situation, Faktor 2: Budgetierung, Faktor 3: sensibler Umgang mit Geld). Die Drei-Faktoren-Lösung bildet die Teilbereiche finanzieller Allgemeinbildung gut ab. Die erklärte Varianz der Faktoren ist zufriedenstellend (39,03 %). Trotz der im Allgemeinen positiven Befunde wurden einige Limitationen deutlich. Während sich die mäßige interne Konsistenz der Skalen, gemessen durch Cronbachs Alpha, bei Werten zwischen ‚573 und ‚754 noch im Einklang mit vielen anderen SJT befindet (s. Metaanalyse von Catano, Brochu und Lamerson 2012), mussten in Bezug auf andere messmethodische Kennzahlen (Trennschärfe, Itemschwierigkeit) Überarbeitungen durchgeführt werden. Grundlagen hierfür waren mehrere Studien:

i. Interviewstudien anlässlich mangelhafter Itemtrennschärfen und -schwierigkeiten

Für die Items 1, 2 und 4 der Situation 2 und für die Items 1 und 3 der Situation 4 wurde eine nicht zufriedenstellende Trennschärfe < ,30 festgestellt. In einer think-aloud-Studie mit 11 Teilnehmenden (vgl. Staub-Kaminsky 2020) wurde deshalb jeweils nur der Itemstamm beibehalten und die Proband:innen wurden offen gefragt, wie sie sich in einer solchen Situation verhalten würden. Dabei konnten Missverständnisse bzw. Unkenntnis der verwendeten Terminologie identifiziert werden: Einige Proband:innen kannten den Begriff ‚Ratenzahlung' nicht, 5 Proband:innen waren sich nicht sicher, was mit ‚Ausgaben' gemeint ist, 2 Proband:innen nahmen an, dass Steuern in den Ausgaben enthalten sind, die meisten Proband:innen kannten den Unterschied zwischen brutto und netto nicht. Die Items wurden auf Basis dieser Studie überarbeitet. Zum Vergleich ist im Folgenden die Situation 4 mit den dazugehörigen Handlungsoptionen in ihrer alten sowie der überarbeiteten Variante abgebildet (s. Abbildung 1 und 2):

Situation 4 (alt):

Sie haben ihre erste Gehaltsüberweisung erhalten und stellen fest, dass Sie weniger überwiesen bekommen haben als ursprünglich im Vertrag festgeschrieben. Dies sorgt dafür, dass Ihre Ausgaben höher als Ihre Einnahmen sind.

Bitte bewerten Sie für jede der nachfolgend beschriebenen Handlungsoptionen, wie wahrscheinlich Sie sich in der beschriebenen Situation genau so verhalten würden.
(1 = sehr unwahrscheinlich, 2 = unwahrscheinlich, 3 = weder noch, 4 = wahrscheinlich, 5 = sehr wahrscheinlich)

	1	2	3	4	5
Ich gehe dem Unterschied auf den Grund und stelle fest, dass ich nicht bedacht habe, dass im Vertrag nur der Bruttobetrag ausgewiesen wird.					
Ich stelle einen geringen Unterschied zwischen erwartetem und überwiesenem Gehalt fest. Daraufhin recherchiere ich im Internet z. B. über einen Online-Gehaltsrechner, informiere mich bei meiner Familie oder Freunden oder suche das direkte Gespräch mit meinem Arbeitgeber, um die Gründe der leichten Abweichung in Erfahrung zu bringen.					
Ich suche nach möglichen Ursachen und finde heraus, dass mir nicht bewusst war, dass neben den Steuerabgaben auch noch Sozialabgaben von meinem Bruttogehalt abgehen.					
Ich stelle einen geringen Unterschied zwischen erwartetem und überwiesenem Gehalt fest. Ich bin verwundert, aber ich unternehme erst einmal nichts.					

Abbildung 1: alte Version der Situation vier und ihre Handlungsoptionen

Situation 4 (neu):

Sie haben einen Minijob und erhalten monatlich €350 auf Ihr Konto ausgezahlt. Da Sie monatlich €300 Ausgaben haben, suchen Sie nach einem besser bezahlten Minijob. Schon nach wenigen Tagen erhalten Sie ein Jobangebot, bei dem Sie laut Vertrag €600 erhalten.

Bitte bewerten Sie für jede der nachfolgend beschriebenen Handlungsoptionen, wie wahrscheinlich Sie sich in der beschriebenen Situation genau so verhalten würden. (1 = sehr unwahrscheinlich, 2 = eher unwahrscheinlich, 3 = eher wahrscheinlich, 4 = sehr wahrscheinlich)

	1	2	3	4
Ich nehmen das Jobangebot an, weil ich die extra €250 monatlich gut gebrauchen kann.				
Ich vermute, dass im Vertrag das Bruttogehalt angegeben wird. Ich informiere mich z. B. über Online-Gehaltsrechner wie viel ich nach Abzug der Steuern und Sozialabzügen bekommen würde und entscheide danach, was ich mache.				
Ich nehme an, dass im Vertrag das Bruttogehalt angegeben wird. Ich denke aber, dass ich nach Abzug der Steuer und Sozialabgaben bestimmt mehr als die vorherigen €350 erhalte und nehme das Jobangebot an.				
Ich denke, dass im Vertrag das Bruttogehalt angegeben wird. Aber da ich bestimmt unter der steuerfreien Grenze bin, würden mir die €600 komplett ausbezahlt. Deshalb nehme ich den Job an.				

Abbildung 2: revidierte Version der Situation vier und ihre Handlungsoptionen

Auch für die Items 2 und 3 der Situation 5 lag eine nicht zufriedenstellende Trennschärfe vor. Mit dieser Situation wird getestet, wie gut Personen der Zielgruppe in der Lage sind, Ausgaben in Abstimmung mit eigenen Bedürfnissen und Möglichkeiten zu planen. Um bessere Antwortalternativen zu entwickeln, wurde in einer Interviewstudie gefragt, wie gut eine solche Abstimmung gelingt und was Gründe dafür sind, dass sich Personen ggf. auch verschulden (vgl. Schill 2021). Aus den Interviewergebnissen wurden 4 neue Antwortalternativen generiert.

Des Weiteren waren viele Items (insgesamt 13) zu leicht. Hier kann soziale Erwünschtheit eine Rolle spielen, da es sich v.a. um solche Items handelt, die recht explizit erwünschtes Verhalten beschreiben (z. B. Geld sparen) oder unerwünschtes als falsch nahelegen (z. B. sich verschulden). Durch soziale Erwünschtheit werden die Testergebnisse verzerrt und deren valide Interpretation verhindert (vgl. Moosbrugger/Brandt 2020, 82). In den bereits genannten think-aloud- und Interviewstudien wurde deshalb nach Anhaltspunkten gesucht, ob und wie Proband:innen die Handlungsmöglichkeiten als erwünscht oder unerwünscht wahrnehmen. Die Antwortalternativen wurden entsprechend überarbeitet. Situation 11 wurde eliminiert, da zur Überarbeitung dieser Items noch umfassendere Analysen notwendig sind. Die übrigen Items decken weiterhin ein ausreichend breites inhaltliches Spektrum ab.

ii. Zwischengruppendesign zur Frage nach dem Einfluss der Instruktion auf das Antwortverhalten der Proband:innen

Laut McDaniel et al. (2014, 184) haben nicht nur der Realitätsbezug, die Komplexität der Items und der Inhalt der Situationen einen Einfluss auf die messmethodische Qualität von SJT, son-

dern auch die Aufgabeninstruktion. Grundsätzlich musste entschieden werden, ob die Itemformulierung Verhalten abfragt, das gezeigt werden sollte (should; was sollte man tun) oder ob gefragt wird, was Testteilnehmer:innen in einer geschilderten Situation tun würden (would; was würden Sie tun). Bei Analyse der einschlägigen Literatur findet sich zu den Vor- und Nachteilen der Versionen kein klares Ergebnis (für eine Übersicht s. McDaniel et al. 2007, 70; Muck 2013, 196). Da die ‚sollte'-Alternative wissensbasiert ist, aber das Ziel darin besteht, mit dem Instrument verhaltensnahe Entscheidungen zu erfassen, wurde bei dem Testeinsatz im Jahr 2016/2017 die zweite Variante (was würden Sie tun) gewählt. Im Zuge des Überarbeitungsprozesses wurde geprüft, ob die beiden Varianten Unterschiede hervorbringen.

Um die Frage potentieller Testergebnisse in Abhängigkeit der Formulierung (was sollte man tun/was würden Sie tun) zu prüfen, wurden die Situationen 1, 3 und 6 bis 10 in einem Zwischengruppendesign eingesetzt (vgl. Großberndt 2020). Die 59 Teilnehmenden (26 männlich, 33 weiblich; Alter M = 16,9, SD = 0,62) wurden den beiden Varianten zufällig zugeordnet. Verglichen wird auf Itemebene. Die Tests zeigen mit wenigen Ausnahmen (Item 3 aus Situation 6 (t = 2,468, p = ,018), 5 aus Situation 7 (t = -2,214, p = ,032) und 4 aus Situation 9 (t = -2,558, p = ,013) keine signifikanten Unterschiede zwischen den beiden Testvarianten. Bei den genannten Items schneiden die Proband:innen mit der ‚should'-Version bei dem dritten Item der Situation 6 besser ab, bei den beiden anderen Items die Proband:innen mit der ‚would'-Version (vgl. Großberndt 2020). Da zwischen den Varianten keine systematischen signifikanten Unterschiede gefunden wurden und da die Variante mit der Abfrage des typischen (eigenen) Verhaltens näher am tatsächlichen Verhalten von Proband:innen liegt, wurde in der 2023er Version des SJT diese Variante eingesetzt.

iii. Literaturrecherchen zur Frage nach der Gestaltung der Likert-Skalen

Wie viele Stufen eine Likert-Skala haben sollte, wird grundsätzlich von verschiedenen Überlegungen abhängig gemacht. Bei einer fünfstufigen Skala wird die mittlere Stufe vermehrt als Ausweichoption genutzt, vor allem, wenn die Testteilnehmenden sich unsicher sind oder ihre Antworten verbergen möchten (vgl. Bühner 2021, 60). Problematisch ist auch, dass diese Mittelkategorie von den Proband:innen oft nicht einheitlich verstanden und somit verschieden interpretiert wird (vgl. Pospeschill 2022, 52). Ausführlichere Analysen zu der Wahl geeigneter Likert-Skalen bei dem SJT Format gibt es bislang keine. In der Literatur finden sich Tests, bei denen eine gerade Stufenanzahl genutzt (z. B. Whetzel/McDaniel 2009) aber auch solche, bei denen eine ungerade Stufenanzahl verwendet wurde (z. B. Clevenger et al. 2001). Weil zu vermuten ist, dass bei der ersten Testdurchführung 2016/2017 Probleme im Zusammenhang mit der ungeraden Stufenanzahl auftraten, wird die Skala, die ursprünglich fünfstufig war (vgl. Wuttke/Aprea 2018, 279), nun auf vier Stufen beschränkt. Die mittlere Stufe (‚weder noch') wird also eliminiert.

Der überarbeitete Test umfasst 10 Situationen. In jeder Situation sind 4 Handlungsoptionen auf einer vierstufigen Skala einzuschätzen. Nur Situation 7 enthält 6 Handlungsoptionen. Insgesamt gibt es folglich 42 Testitems. An der inhaltlichen Ausrichtung des Tests hat sich nichts geän-

dert, d. h. der SJT erfasst nach wie vor das Planen und Verwalten alltäglicher Geldangelegenheiten und deckt dabei 3 Subfacetten des individuell-kognitiven Kompetenzbereichs ab (Bewertung der eigenen Einnahmen, Planung der Ausgaben in Übereinstimmung mit den eigenen Bedürfnissen und Möglichkeiten/Einnahmen, Aufstellen eines Budgets).

3.2 Stichprobe

Mit der überarbeiteten Version wurden im Zeitraum von Mai 2023 bis April 2024 zum einen Schüler:innen aus dem berufsbildenden Bereich befragt. Dabei wurde darauf geachtet, verschiedene Schulformen abzudecken. Zum anderen wurden aus der Zielgruppe der jungen Erwachsenen auch Studierende erhoben. Bei der Stichprobe handelt es sich um eine Gelegenheitsstichprobe (vgl. Döring 2023, 307). Sie sollte sowohl Personen mit finanzieller Vorbildung beinhalten als auch solche, die bislang auf ihrem Bildungsweg wenig Berührung mit finanzbezogenen Themen hatten.[3] Die Größe der Stichprobe sollte im Vergleich zur ersten Testdurchführung im Jahr 2016/2017 außerdem noch einmal erhöht werden. Ferner wurde angestrebt, Proband:innen aus zwei verschiedenen Bundesländern (Hessen und Sachsen) zu befragen. Die Testdurchführung erfolgte in zwei Etappen. Während in der ersten Hälfte des Testzeitraumes vor allem Proband:innen in Sachsen befragt wurden (vgl. Wagner 2023), fand die Befragung in Hessen in der zweiten Hälfte des Testzeitraumes statt.

Der Test wurde im Paper-Pencil-Format eingesetzt, um die Teilnehmenden durch den persönlichen Kontakt mit der Testleitung zu motivieren (vgl. Reinders/Post 2022, 181), sodass diese bei der Bearbeitung konzentrierter sind und die Rücklaufquoten höher ausfallen (vgl. Kraus/Kreitenweis 2020, 232). Die Befragung nahm ca. 45 Minuten in Anspruch. Der Fragebogen, der den Proband:innen vorgelegt wurde, enthielt neben dem SJT einige sozio-biografische Angaben, sowie den Fragebogen zum Kaufverhalten von Ray und Najman (1986). Dieser wird als externes Kriterium im Rahmen des Validierungsprozesses (vgl. AERA/APA/NCME, 2014) hinzugezogen. Insgesamt konnte der SJT mit $N = 885$ Personen aus acht verschiedenen Schulen und drei unterschiedlichen Studiengängen durchgeführt werden. Von den befragten Personen leben ca. zwei Drittel in Hessen und ein Drittel in Sachsen. Nähere Informationen zur Zusammensetzung der Stichprobe finden sich in der Tabelle 2.

[3] Es wurden Schüler:innen aus kaufmännischen und nicht-kaufmännischen Klassen des berufsbildenden Bereichs sowie Studierende aus wirtschaftsbezogenen Studiengängen (Bachelor- und Masterstudium Wirtschaftspädagogik sowie Bachelorstudium Wirtschaftswissenschaften) befragt. Eine Testdurchführung mit Studierenden aus anderen Studiengängen steht noch aus.

Tabelle 2: Beschreibung der Stichprobe

		N = 885
Bildungsweg	Fachoberschule	78
	Berufliches Gymnasium	70
	Berufsausbildung	588
	BÜA (Berufsfachschule zum Übergang in Ausbildung)	9
	Fachschule	13
	Studium (Wirtschaftswissenschaften, Wirtschaftspädagogik)	127
Bundesland	Sachsen 36,00 %, Hessen 64,00 %	
Geschlecht	männlich 47,50 %, weiblich 52,20 %, divers 0,30 %	
Beste Sprache	Deutsch 71,30 %, Deutsch + andere Sprache 22,60 %, andere Sprache 6,10 %	
Alter	*M* = 20,71 Jahre, *SD* = 3,40	

3.3 Überprüfung des Tests

3.3.1 Dimensionalität des Tests

Um die revidierte Version des SJT messmethodisch zu bewerten, soll zunächst die Dimensionalität des Tests festgestellt werden. Dafür wird eine Explorative Faktorenanalyse (EFA) durchgeführt, damit auch gänzlich neue Beziehungsstrukturen unter den Items identifizieren werden können Diese sind nicht auszuschließen, da einzelne Items und auch ganze Situationen gestrichen und überarbeitet wurden, sodass eine stark modifizierte Version des Instruments bei der zweiten Testung 2023 und 2024 vorliegt. Nach der Überprüfung der Voraussetzungen (Kaiser-Meyer-Olkin-Koeffizient (KMO), Bartlett-Test, Prüfen der Kommunalitäten, vgl. Pospeschill 2022, 208f.) sind grundsätzlich drei Entscheidungen zu treffen: die Wahl der Methode, des Abbruchkriteriums und der Faktorenrotation (vgl. Brandt 2020, 578). In Anlehnung an das Vorgehen von Wuttke und Aprea (2018), wird im ersten Schritt eine Hauptkomponentenanalyse (Principal Component Analysis, PCA) zur Extraktion der Faktoren durchgeführt, die der Dimensionsreduktion dient und möglichst viel Varianz der beobachteten Werte erklären soll.

Zur Bestimmung der Abbruchkriterien wird die Parallelanalyse nach O'Connor (2000) gewählt, bei der Eigenwerte aus Zufallsdatensätzen extrahiert werden. Die Zufallsdatensätze gleichen dem tatsächlichen Datensatz in Bezug auf die Anzahl an Variablen sowie in Bezug auf die Stichprobengröße. Berechnet wird das 95. Perzentil aus der Verteilung der Eigenwerte. Die Faktoren werden dann ermittelt, indem die Eigenwerte aus dem tatsächlichen Datensatz mit den Eigenwerten aus den Zufallsdatensätzen verglichen werden (vgl. O'Connor 2000, 397). Zusätzlich wird der Scree-Test mit der Varimax-Rotation durchgeführt (vgl. Pospeschill 2022, 206f.).

Zur Entscheidung, welche Variable auf einen Faktor lädt, nutzen wir den Marker Index von Gallucci und Perugini (2007). Er berücksichtigt neben den absoluten Werten der Primärladungen auch das Verhältnis zwischen den Primär- und Sekundärladungen. Die berechneten Indizes sollen > ,40 sein (vgl. Gallucci/Perugini 2007, 7f.).

3.3.2 Itemanalyse

Um Trennschärfe und Itemschwierigkeit zu bestimmen, werden Itemanalysen durchgeführt. Während einige Quellen die untere Grenze bei der Itemtrennschärfe auf ,40 (vgl. Kelava/Moosbrugger 2020, 155) oder ,30 (vgl. Lienert/Raatz 1998, 106) setzen, wird im Financial Literacy-Bereich durchaus auch ,20 als Cut-Off-Wert genutzt (z. B. beim TFL: Walstad/Rebeck 2017, 117 oder dem BFT/TFK: Walstad/Rebeck 2018, 261). Items, die unterhalb dieser Grenze liegen ($r_{it} <$,20), gelten als nicht trennscharf und werden von den weiteren Analysen ausgeschlossen.

In Bezug auf die Itemschwierigkeit gilt, dass die Items ähnlich schwierig sein und einen Wert im mittleren Bereich aufweisen sollten (vgl. Pospeschill 2022, 85). Sie dürfen also weder zu schwer noch zu leicht sein. Zur Festlegung dieser Grenzen orientieren wir uns an Wuttke und Aprea (2018, 281) ($p_i <$,20: Items sind zu schwer; ,90 $< p_i \leq 1$: Items sind zu leicht).

3.3.3 Reliabilität

Im Rahmen der Reliabilitätsanalyse wird die Zuverlässigkeit der Skalen anhand von Cronbachs Alpha überprüft (vgl. Schnell/Hill/Esser 2018, 133f.). Angestrebt wird in der Regel ein Cronbachs Alpha von rund ,80. Werte, die deutlich kleiner sind, weisen darauf hin, dass die Skalen nicht reliabel sind. Werte, die größer als ,80 sind, erscheinen jedoch auch nicht erstrebenswert, da dies ein Hinweis darauf sein kann, dass Items enthalten sind, die Redundanzen aufweisen (vgl. Streiner 2003, 103).[4]

3.3.4 Validierungsstandards

Im Rahmen des Validierungsprozesses werden einzelne, bewusst ausgewählte Belege dafür gesammelt, ob eine valide Testwertinterpretation möglich ist. Entsprechend der Standards for Educational and Psychological Testing der AERA, APA und NCME (2014, 14ff.) basieren diese Belege auf dem Testinhalt, den Antwortprozessen, der internen Struktur, der Beziehung zu anderen Variablen oder den Konsequenzen der Testung. In diesem Beitrag soll der Fokus auf einer der fünf Hauptquellen für Validitätsnachweise liegen: die Beziehung zu einer anderen Variable. Das Hinzuziehen externer Kriterien bezeichnen Schmitt und Chan (2014, 141) als das wahrscheinlich am häufigsten genutzte Vorgehen, das im Rahmen der Validierung von SJT gewählt wird. Dabei wird ein weiteres Instrument verwendet, welches ein ähnliches Konstrukt wie der SJT misst. Der Fragebogen von Ray und Najman (1986) bietet sich in unserem Fall an, weil er das Kaufverhalten erfasst, das mit der im SJT gemessen FL zusammenhängen sollte. Die Proband:innen treffen auch hier ihre Entscheidungen, in dem sie jedes der Items auf einer vierstufigen Likert-Skala einschätzen. Überprüft werden soll bei dem Vorgehen das Vorliegen konvergenter Evidenz (vgl. AERA/APA/NCME 2014, 16f.). Es ist zu vermuten, dass bei Proband:innen, die ein rationales Kaufverhalten vorweisen, auch eine hohe FL festzustellen ist, da der SJT den Anspruch hat, FL verhaltensnah zu messen.

[4] Der vorliegende Beitrag ist im Zuge der Masterarbeit von Elisa Wagner entstanden. Er fußt auf der klassischen Testtheorie. Die Schätzung einer EAP-Reliabilität wurde deshalb bislang noch nicht unternommen, kann aber in dem Folgeprojekt (s. Ausblick) anvisiert werden.

4 Befunde

4.1 Dimensionalität

Mit einem KMO von ,851, einem signifikanten Bartlett-Test und Kommunalitäten über ,50, sind die Voraussetzungen zur Durchführung einer Faktorenanalyse erfüllt. Die Parallelanalyse hat ergeben, dass zunächst sieben Faktoren extrahiert werden sollten, wohingegen der Scree-Test kein eindeutiges Ergebnis präsentiert (s. Abbildung 3).

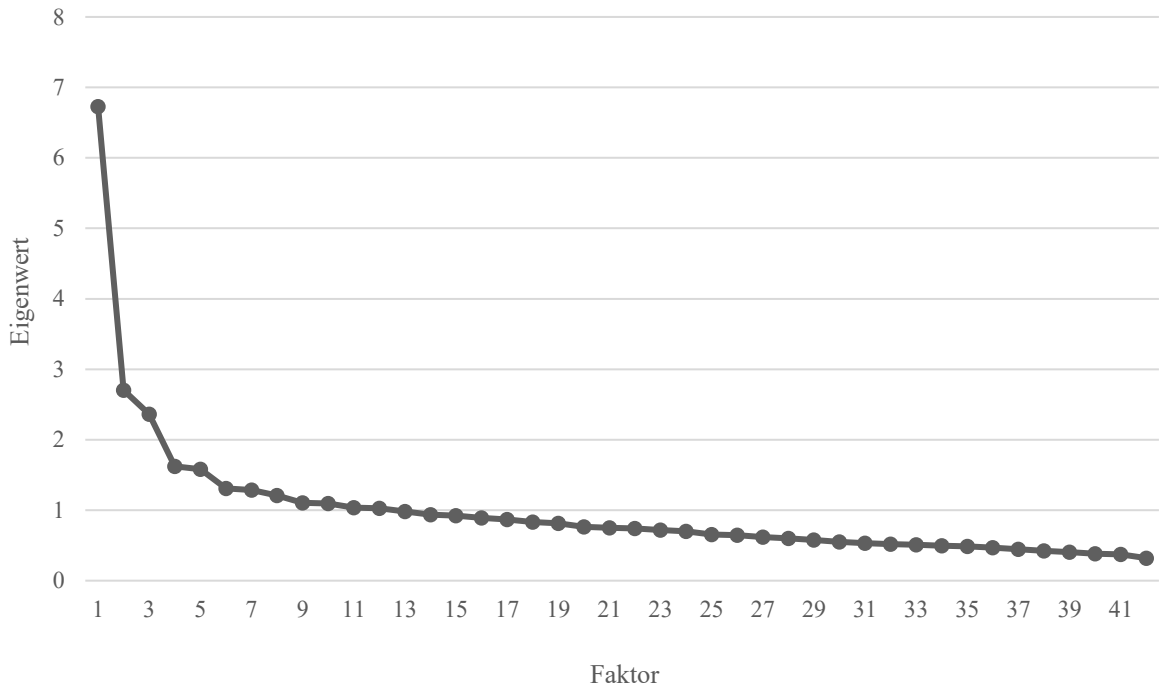

Abbildung 3: Scree-Test

Da bei Durchführung der Hauptkomponentenanalyse mit Varimax-Rotation mehrmals beobachtet wurde, dass auf die jeweils letzten Faktoren nur ein oder zwei Items laden (die Mindestanzahl jedoch üblicherweise bei drei liegt: vgl. Izquierdo/Olea/Abad 2014, 399; MacCallum et al. 1999, 96), musste die Zahl der Faktoren weiter reduziert werden. Letztlich ist eine Fünf-Faktoren-Lösung zu identifizieren. Diese erklärt insgesamt 50,23 % der Varianz. Von den ursprünglich 42 Items bleiben nach Untersuchung der Dimensionalität noch 21 Items übrig, die sich wie folgt auf die fünf Faktoren aufteilen (s. Tabelle 3).

Tabelle 3: Rotierte Komponentenmatrix

Item	Inhalt	Komponenten				
		1	2	3	4	5
Sit1.3	Leihen von zusätzlichem Geld, um sich dann ein Smartphone kaufen zu können	**,531**	-,115	,197	-,040	-,268
Sit1.4	Sparen des eigenen Geldes, um sich später ein Smartphone kaufen zu können	**,560**	-,051	,255	-,179	-,034
Sit5.1	Sofortiges Kaufen eines neuen Smartphones, obwohl das alte noch funktionstüchtig ist	**,609**	,007	,060	-,012	,064
Sit5.4	Kaufen zwei neuer Güter (Smartphone & Laptop), ohne die finanziellen Möglichkeiten dafür zu haben	**,678**	,045	,104	,022	-,021
Sit6.2	Mieten einer Wohnung, obwohl dafür die finan-ziellen Mittel fehlen	**,702**	,051	,022	,108	,050
Sit6.4	Analysieren der Kosten für eine neue Wohnung im Vergleich zum Einkommen	**,650**	,091	-,039	,168	-,043
Sit7.1	Konto überziehen, um sich einen Laptop kaufen zu können	**,672**	,146	-,102	,075	,141
Sit9.3	Keinen eigenen Budgetplan mit Einnahmen und Ausgaben aufstellen	**,563**	,175	,098	,121	,088
Sit8.1	Einen neuen Budgetplan (schriftlich) erstellen bei Aufnahme einer Ausbildung	,081	**,819**	,173	-,009	,084
Sit8.2	Einen neuen Budgetplan (im Kopf) erstellen bei Aufnahme einer Ausbildung	,054	**,743**	,068	-,028	-,278
Sit8.4	Keinen Budgetplan erstellen bei Aufnahme einer Ausbildung	,117	**,578**	,053	,268	,044
Sit9.2	Einen eigenen Budgetplan mit Einnahmen und Ausgaben aufstellen	,049	**,705**	,251	-,061	,122
Sit3.1	Kein Beurteilen der Folgen eines Anstiegs der Krankenversicherungskosten auf das Budget	,166	,138	**,731**	,203	-,030
Sit3.3	Teilweise Beurteilen der Folgen eines Anstiegs der Krankenversicherungskosten auf das Budget	,157	,149	**,733**	,133	-,078
Sit3.4	Beurteilen der Folgen eines Anstiegs der Krankenversicherungskosten auf das Budget	,019	,228	**,772**	-,039	,054
Sit4.1	600€- statt 350€-Job: Jobannahme aufgrund von 250€ mehr Gehalt pro Monat	-,040	,017	,218	**,744**	,071
Sit4.3	600€- statt 350€-Job: Jobannahme, weil vermutlich nach Abziehen von Steuer usw. mehr Gehalt bleibt	,019	,080	,021	**,653**	-,209
Sit4.4	600€- statt 350€-Job: Jobannahme, weil Gehalt von 600€ vermutlich unterhalb der Minijob-Grenze liegt	,228	-,009	,011	**,642**	,081
Sit2.2	Von 100€ (brutto) die Hälfte auf ein gut verzinstes Konto anlegen für eine Investition in einem Jahr	-,053	,035	-,110	,071	**-,581**
Sit5.2	Sofortiges Kaufen eines Laptops, da dieser für die Uni gebraucht wird (statt abwarten bis Preis sinkt)	-,078	,084	,012	,056	**,637**
Sit10.4	Entscheiden, dass ein Anpassen des Budgetplans nach einmaliger hoher Ausgabe nicht notwendig ist	,118	-,044	-,189	-,026	**,619**
Eigenwerte		3,99	2,33	1,55	1,43	1,25
Erklärte Varianz (in %)		15,54	10,69	9,54	7,77	6,69

Anmerkung.
Hauptkomponentenanalyse (PCA) mit Varimax-Rotation. Die Primärladungen sind **hervorgehoben**. Die Entscheidung, welches Item eindeutig auf einen Faktor lädt, wurde mit dem Marker-Index von Gallucci und Perugini (2007) gefällt.

4.2 Itemanalyse

Bei der Itemanalyse konnten 2 Items als zu leicht identifiziert werden (Item 2 aus Situation 6 sowie Item 1 aus Situation 7). Keins der betrachteten 21 Items ist zu schwer. Hinsichtlich der Trennschärfe weisen alle Items, die auf den fünften Faktor laden, unzureichende Werte auf. Dieser wird deshalb von den weiteren Analysen vollständig ausgeschlossen. In Summe bleiben 16 Items übrig. Die Werte können Tabelle 4 entnommen werden.

Tabelle 4: Itemanalyse

	Item	Mittelwert	Standard-abweichung	Itemschwierig-keit p_i	Itemtrenn-schärfe r_{it}
Faktor 1	Sit1.3	2,53	0,814	,843	,398
	Sit1.4	2,60	0,810	,867	,421
	Sit5.1	2,69	0,708	,897	,461
	Sit5.4	2,70	0,703	,900	,542
	Sit6.2*	2,79	0,570	,930	,560
	Sit6.4	2,63	0,744	,877	,502
	Sit7.1*	2,87	0,508	,957	,511
	Sit9.3	2,52	0,783	,840	,453
Faktor 2	Sit8.1	1,84	1,064	,613	,651
	Sit8.2	1,55	1,055	,517	,491
	Sit8.4	2,11	0,958	,703	,374
	Sit9.2	1,48	1,042	,493	,524
Faktor 3	Sit3.1	1,93	0,995	,643	,541
	Sit3.3	1,97	0,989	,657	,551
	Sit3.4	1,86	1,094	,620	,542
Faktor 4	Sit4.1	1,43	1,150	,477	,399
	Sit4.3	1,51	0,976	,503	,310
	Sit4.4	1,94	1,064	,647	,292
Faktor 5	Sit2.2*	1,32	0,985	,440	-,183
	Sit5.2*	1,86	0,912	,620	,009
	Sit10.4*	1,76	0,948	,587	-,016

Anmerkung.
Die mit * markierten Items werden ausgeschlossen, wenn $p_i < ,20$ oder $,90 < p_i < 1$. Sie werden ferner ausgeschlossen bei $r_{it} < ,20$.

Inhaltlich können die 4 verbleibenden Faktoren wie folgt interpretiert werden: Fähigkeit zum Belohnungsaufschub (Faktor 1), Budgetierung (Faktor 2), Einschätzen der eigenen Einnahmen (Faktor 3) und Einschätzen des Netto-/Bruttogehaltes (Faktor 4). Die Faktoren korrelieren alle positiv miteinander ($\rho s = ,137$ bis $,370$ mit jeweils $p \leq ,01$).[5] In der Entwicklung und Überarbeitung des Tests wurde bereits damit gerechnet, dass einige Items in den Analysen ausgeschlossen werden, weshalb absichtlich etwas mehr Items konstruiert wurden als anfänglich not-

[5] Es wurden Korrelationen nach Spearman durchgeführt, da die Daten zwar intervallskaliert, aber nicht normalverteilt sind. Die Annahme der Normalverteilung spielt auch bei der Faktorenanalyse eine wichtige Rolle (vgl. Janssen/Laatz 2017, 578). Sie wurde aber trotzdem durchgeführt, weil artverwandte Methoden (z. B. die Clusteranalyse) im Vergleich zur Faktorenanalyse einige Schwächen vorweisen (vgl. Überla 1971, 307).

wendig erschien. Die Anzahl an Items ist nun allerdings so gering, dass sich die Interpretationsgrundlage des Tests deutlich minimiert. Die Interpretation der Vier-Faktoren-Lösung ist somit erschwert.

4.3 Reliabilitätsanalyse

Die Reliabilitäten der Skalen sind nur bedingt zufriedenstellend (α = ,708 für Faktor 1; α = ,720 für Faktor 2; α = ,723 für Faktor 3; α = ,517 für Faktor 4). Insbesondere Cronbachs Alpha beim vierten Faktor zeigt noch einmal, dass eine weitere Überarbeitung des Tests von Nöten ist. Die Analyse hat außerdem ergeben, dass das Ausschließen weiterer Items keine Verbesserung der Reliabilitätswerte hervorbringen würde. Für die Korrelation, die im Rahmen des Validierungsprozesses noch durchgeführt wird, bleibt es deshalb bei 16 Items. Davon laden 6 Items auf den ersten, 4 Items auf den zweiten, 3 Items auf den dritten und 3 Items auf den vierten Faktor.

4.4 Validierung mit einem Außenkriterium

Bei der Korrelation nach Spearman ist festzustellen, dass statistisch signifikante Zusammenhänge zwischen allen Skalen des SJT und dem Kaufverhalten bestehen (ρ = ,411 für Faktor 1; ρ = ,277 für Faktor 2; ρ = ,331 für Faktor 3; ρ = ,198 für Faktor 4; wobei $p \leq$,01). Die Befunde dienen aus diesem Grund als Beleg für eine valide Testwertinterpretation. Gleichwohl ist zu betonen, dass die Korrelationskoeffizienten verhältnismäßig klein sind und die ermittelten, positiven Zusammenhänge eher als schwach gelten. Allerdings wäre bei zu hohen Korrelationen auch zu hinterfragen, ob die beiden Instrumente noch eigenständige Konstrukte erfassen. Eine sehr hohe Korrelation dürfte nur vorliegen, wenn die zwei Tests tatsächlich dasselbe messen (vgl. Bühner 2021, 35). Aus diesem Grund werden die Befunde zwischen ,40 und ,20 als zufriedenstellend eingestuft.

5 Diskussion und Ausblick

Der vorliegende Beitrag beschreibt die Überarbeitung eines SJT zur Erfassung der FL. Es wurden auf der Basis der Befunde der ersten Version Überarbeitungsschritte unternommen, um mit Blick auf die Testgüte wünschenswertere Befunde zu erzielen. Zu dem überarbeiteten Test wurden in dem Beitrag ausgewählte Befunde aus messmethodischer Sicht und aus der Perspektive der Validierung vorgestellt. Im Verhältnis zur alten Testversion (2016/2017) konnten nun vier statt drei Dimensionen ermittelt werden. Die inhaltliche Interpretation dieser Dimensionen – insbesondere der vierten – ist aufgrund der geringen Anzahl an Items erschwert, aber noch möglich. Des Weiteren gibt es nach wie vor Items, die in der Itemanalyse schlecht abschneiden und dabei nicht zufriedenstellende Trennschärfen oder Schwierigkeiten vorweisen. Die Reliabilitäten sind nur zum Teil zufriedenstellend und weisen ähnliche Werte wie in der alten Testversion auf. Die statistisch signifikante Korrelation mit dem externen Kriterium ‚Kaufverhalten', gemessen durch den Fragebogen von Ray und Najman (1986), ist als Beleg für eine valide Testwertinterpretation zu deuten.

Die Befunde sind so zu interpretieren, dass keine eindeutige Verbesserung in der messmethodischen Qualität des Instruments erzielt werden konnte (s. den Vergleich der beiden Versionen in Tabelle 5). Während die Items der Situation 4 nach der Faktorenanalyse noch Berücksichtigung finden und gute Trennschärfen und Schwierigkeiten aufweisen, ist dies beispielsweise für die Situation 2, bei der ebenfalls umfassende Überarbeitungen durchgeführt wurden, nicht der Fall. Auffallend ist zudem, dass bereits nach Überprüfen der Dimensionalität eine große Anzahl an Items ausgeschlossen wird. Insgesamt ist die Anzahl an verbleibenden Items im direkten Vergleich zur alten Testversion geringer, sodass eine adäquate Messung des zugrundeliegenden latenten Merkmals nur eingeschränkt möglich ist (vgl. MacCallum et al. 1999, 96f.). Die Gegenüberstellung der zwei Testversionen verdeutlicht ebenfalls, dass es nur wenige Überschneidungen in der Faktorenstruktur gibt. Während Faktor 2 der alten Version und Faktor 2 und 3 der neuen Testversion Überschneidungen hinsichtlich der Items und somit auch in Bezug auf die inhaltliche Interpretation (Budgetierung bzw. Einschätzen der eigenen Einnahmen) zeigen, sind die anderen Faktoren nur schwer miteinander in Beziehung zu setzen. Insgesamt weisen sowohl die alte als auch die neue Testversion Schwächen auf. Weitere Überarbeitungsschritte sind deshalb notwendig.

Tabelle 5: Vergleich der Testgüte (alte vs. revidierte Version des SJT)

	Alte Testversion	**Revidierte Testversion**
Dimensionalität	Scree-Test; PCA mit Varimax-Rotation; Ladungen > ,30 und keine Doppel-ladungen: 1) Überblick/Kontrolle der eigenen finanziellen Situation 2) Budgetierung 3) sensibler Umgang mit Geld	Scree-Test und Parallelanalyse; PCA mit Varimax-Rotation; Marker-Index: 1) Fähigkeit zum Belohnungsaufschub 2) Budgetierung 3) Einschätzen der eigenen Einnahmen 4) Einschätzen des Brutto- bzw. Nettogehaltes
Itemanalyse*	Einige Items zu leicht oder nicht trennscharf	Einige Items zu leicht oder nicht trennscharf
Reliabilitätsanalyse	Cronbachs Alpha: α = ,754 (für Faktor 1) α = ,573 (für Faktor 2) α = ,691 (für Faktor 3)	Cronbachs Alpha: α = ,708 (für Faktor 1) α = ,720 (für Faktor 2) α = ,723 (für Faktor 3) α = ,517 (für Faktor 4)
Validierung	Korrelation nach Pearson; Analyse mit den verbleibenden 26 von 46 Items: r = ,568 (für Faktor 1) r = ,465 (für Faktor 2) r = ,336 (für Faktor 3) (mit jeweils $p \leq$,01)	Korrelation nach Spearman; Analyse mit den verbleibenden 16 von 42 Items: ρ = ,411 (für Faktor 1) ρ = ,277 (für Faktor 2) ρ = ,331 (für Faktor 3) ρ = ,198 (für Faktor 4) (mit jeweils $p \leq$,01)

Anmerkung.
*In Wuttke und Aprea (2018) wurde zunächst die Itemanalyse und danach die Faktorenanalyse durchgeführt. Ein direkter Vergleich zwischen der Anzahl an Items, die zu leicht ist bzw. mangelhafte Trennschärfen vorweist, ist deshalb nicht zielführend.

Aktuell wird sich auf zwei Ebenen der Herausforderung der nicht zufriedenstellenden Gütekriterien für den SJT gestellt. Die erste Ebene ist eine messmethodische. Auf Basis von Rückmeldungen auf wissenschaftlichen Konferenzen werden im nächsten Schritt Modelle mit Hilfe der Item Response Theorie (vgl. Embretson/Reise 2000; Rost 2004) geprüft. Hier erweist sich das Aufgabenformat über Likert-Skalen und die Mehrdimensionalität des Tests allerdings als herausfordernd, weshalb spezielle multidimensionale IRT Modelle ausgewählt werden müssen. Die zweite Ebene ist eine substantielle Überarbeitung des Testinstruments im Rahmen eines in 2025 startenden Drittmittelprojektes (EVerFit: Entwicklung und Validierung eines technologiebasierten Assessments zur Diagnose handlungsnaher Finanzkompetenz; Förderlinie „Forschung zu finanzieller Bildung“). Geplant ist ein computerbasierter SJT, der ein breites Spektrum an Inhaltsbereichen abdecken soll. Hiervon erwarten wir substantielle Fortschritte. Auch zukünftig liegt der Fokus somit auf der Entwicklung und Validierung geeigneter Messinstrumente. Erst im Anschluss daran können Maßnahmen abgeleitet werden, die junge Erwachsene bei der Bewältigung finanzbezogener Herausforderungen unterstützen. Die Notwendigkeit für solche (Bildungs-)Maßnahmen ergibt sich insbesondere aus der mangelhaften curricularen Verankerung wirtschafts- und finanzbezogener Themen sowie die damit zusammenhängende Lehrkräftebildung (vgl. Schuler/Brahm 2021; Flossbach von Storch Stiftung/IÖB 2021).

Nicht zuletzt ist ein skeptischer Blick auf die Testökonomie (vgl. Bühner 2021, 634) zu werfen. Sicher könnte man argumentieren, dass SJT aufgrund der umfassenden Testzeit nur bedingt ökonomisch sind und dass deshalb manche Bereiche von FL mit anderen Instrumenten (z. B. klassischen Wissenstests) getestet werden könnten. Dennoch bleibt zu beachten: wenn man handlungsnah testen möchte, bleiben SJT sicherlich eine der besten (wenn nicht die beste) Alternative. Zudem muss auch nicht bei jeder Testung und jeder Zielgruppe das gesamte Spektrum der acht Facetten abgeprüft werden, sondern es ist möglich – wenn für alle Facetten gute Tests vorliegen – je nach Diagnosebedarf Teilfacetten einzusetzen.

Literatur

Amagir, A./Wilschut, A./Groot, W. (2018): The Relation between Financial Knowledge, Attitudes towards Money, Financial Self-Efficacy, and Financial Behavior among High School Students in the Netherlands. In: Empirische Pädagogik, 32, H. 3-4, 387-400.

American Educational Research Association (AERA)/American Psychological Association (APA)/National Council of Measurement in Education (NCME) (2014): Standards for Educational and Psychological Testing. Washington.

Aprea, C. (2012): Messung der Befähigung zum Umgang mit Geld und Finanzthemen: Ausgewählte Instrumente und alternative diagnostische Zugänge. In: bwp@ Berufs- und Wirtschaftspädagogik – online, Ausgabe 22, 1-21. Online: www.bwpat.de/ausgabe22/aprea_bwpat22.pdf (27.06.2024).

Aprea, C./Wuttke, E. (2016): Financial Literacy of adolescent and young adults: Setting the course for a competence-oriented assessment approach. In: Aprea, C./Wuttke, E./Breuer, K./Keng, N. K./Davies, P./Greimel-Fuhrmann, B./Lopus, J. (Hrsg.): International handbook of Financial Literacy. Singapur, 397-414.

Aprea, C./Wuttke, E./Leumann, S./Heumann, M. (2015): Kompetenzfacetten von Financial Literacy: Sichtweisen verschiedener Akteure. In: Seifried, J./Seeber, S./Ziegler, B. (Hrsg.): Jahrbuch der berufs- und wirtschaftspädagogischen Forschung 2015. Berlin/Toronto, 11-22.

Atkinson, A./Messy, F. (2012): Measuring Financial Literacy: Results of the OECD/International Network on Financial Education (INFE) Pilot Study. OECD Working Papers on Finance, Insurance and Private Pensions, 15. Paris.

BMF/BMBF (2023): Eckpunkte für finanzielle Bildung. März 2023. Online: https://www.bundesfinanzministerium.de/Content/DE/Downloads/Internationales-Finanzmarkt/eckpunkte-fuer-finanzielle-bildung.pdf?__blob=publicationFile&v=8 (27.06.2024).

Döring, N. (2023): Stichprobenziehung. In: Döring, N. (Hrsg.): Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften. 6. Aufl. Berlin, 293-320.

Brandt, H. (2020): Exploratorische Faktorenanalyse (EFA). In Moosbrugger, H./Kelava, A. (Hrsg.): Testtheorie und Fragebogenkonstruktion. 3. Aufl. Berlin, 575-614.

Breuer, K. (2016): Assessment of Financial Literacy. In: Aprea, C./Wuttke, E./Breuer, K./ Keng, N.K./Davies, P./Greimel-Fuhrmann, B./Lopus, J. (Hrsg.): International handbook of Financial Literacy. Singapur, 381-382.

Bucher-Koenen, T./Knebel, C. (2021): Finanzwissen und Finanzbildung in Deutschland – Was wissen wir eigentlich? In: Vierteljahrshefte zur Wirtschaftsforschung, 90, H. 1, 11-32.

Bühner, M. (2021): Einführung in die Test- und Fragebogenkonstruktion. 4. Aufl. München.

Carmel, E./Leiser, D./Spivak, A. (2020): The Arrested Deployment Model of Financial Literacy. In: Zaleskiewicz, T./Traczyk, J. (Hrsg.): Psychological Perspectives on Financial Decision Making. New York, 89-105.

Catano, V. M./Brochu, A./Lamerson, C. D. (2012): Assessing the reliability of situational judgment tests used in high-stakes situations. In: International Journal of Selections and Assessment, 20, 333-346.

Clevenger, J./Pereira, G. M./Wiechmann, D./Schmitt, N./Harvey, V. S. (2001): Incremental validity of situational judgment tests. In: Journal of Applied Psychology, 86, H. 3, 410-417.

Council for Economic Education (CEE) (2013): National Standards for Financial Literacy. New York.

Embretson, S. E./Reise, S. P. (2000): Item response theory for psychologists. Mahwah.

Flossbach von Storch Stiftung/IÖB (2021): Die OeBiX-Studie. Zum Stand der ökonomischen Bildung in Deutschland. Kernergebnisse. Online: https://www.flossbachvonstorch-stiftung.de/media/pages/downloadcenter/e03bf1f92a-1709564270/oebix-studie-kernergebnisse.pdf (07.08.2024).

Förster, M./Happ, R./Molerov, D. (2017): Using the U.S. Test of Financial Literacy in Germany – Adaptation and Validation. In: The Journal of Economic Education, 48, H. 2, 123-135. https://doi.org/10.1080/00220485.2017.1285737.

Gallucci, M./Perugini, M. (2007): The Marker Index: A new method of selection of marker variables in factor analysis. TPM-Testing, In: Psychometrics, Methodology in Applied Psychology, 14, H. 1, 3-25.

Großberndt, T. (2020): Financial Literacy in der gymnasialen Oberstufe – eine empirische Studie. Unveröffentlichte Bachelorarbeit an der Goethe Universität Frankfurt.

Gruber, H./Renkl, A. (2000): Die Kluft zwischen Wissen und Handeln: Das Problem des trägen Wissens. In Neuweg, G.H. (Hrsg.): Wissen – Können – Reflexion. Ausgewählte Verhältnisbestimmungen. Innsbruck, 155-174.

Happ, R./Förster, M. (2019): The relationship between migration background and knowledge and understanding of personal finance of young adults in Germany. In: International Review of Economics Education, 30, 1-14. https://doi.org/10.1016/j.iree.2018.06.003.

Happ, R./Hahn, J./Jang, K./Rüter, I. (2022): Financial knowledge of university students in Korea and Germany. In: Research in Comparative and International Education, 17, H. 2, 301-327. https://doi.org/10.1177/17454999221086357.

Huston, S.J. (2010): Measuring Financial Literacy. In: The Journal of Consumer Affairs, 44 H. 2, 296-316.

Izquierdo, I./Olea, J./Abad, F.J. (2014): Exploratory factor analysis in validation studies: Uses and recommendations. In: Psicothema, 26, H. 3, 395-400. http://dx.doi.org/10.7334/psicothema2013.349.

Janssen, J./Laatz, W. (2017): Statistische Datenanalyse mit SPSS. Eine anwendungsorientierte Einführung in das Basissystem und das Modul exakte Tests. 9. Aufl. Berlin.

Kahmann, J. (2014): Entwicklung und Validierung eines Situational Judgement Tests (SJT) zur Erfassung sozialer Kompetenzen von Studienplatzbewerbern und -interessenten der Human- und Zahnmedizin. Dissertation, Ruprecht-Karls-Universität Heidelberg: Heidelberg.

Kelava, A./Moosbrugger, H. (2020): Deskriptivstatistische Itemanalyse und Testwertbestimmung. In: Moosbrugger, H./Kelava, A. (Hrsg.): Testtheorie und Fragebogenkonstruktion. 3. Aufl. Berlin, 67-90.

Klapper, L./Lusardi, A./Van Oudheusden, P. (2015): Financial literacy around the world. World Bank. Insights from the Standard and Poor's Rating Services Global Financial Literacy Survey. World Bank.

Kraus, R./Kreitenweis, T. (2020): Führung messen. Inklusive Toolbox mit Messinstrumenten und Fragebögen. Berlin.

Lienert, G.A./Raatz, U. (1998): Testaufbau und Testanalyse. 6. Aufl. Weinheim.

Lusardi, A. (2019): Financial literacy and the need for financial education: evidence and implications. In: Swiss Journal of Economics and Statistics, 155, H. 1, 1-8. https://doi.org/10.1186/s41937-019-0027-5.

Lusardi, A./Mitchell, O.S. (2008): “Planning and Financial Literacy: How Do Women Fare?” In: American Economic Review, 98, H. 2, 413-417. https://doi.org/10.1257/aer.98.2.413.

Lusardi, A./Mitchell, O.S. (2011): Financial literacy around the world: An overview. In: Journal of Pension Economics and Finance, 10, 497-508. https://doi.org/10.1017/S1474747211000448.

Lusardi, A./Mitchell, O.S./Curto, V. (2010): Financial Literacy among the Young. In: The Journal of Consumer Affairs, 44, H. 2, 358-380. https://doi.org/10.1111/j.1745-6606.2010.01173.x.

MacCallum, R.C./Widaman, K.F./Zhang, S./Hong, S. (1999): Sample Size in Factor Analysis. In: Psychological Methods, 4, H. 1, 84-99. https://doi.org/10.1037/1082-989X.4.1.84.

Mandl, H./Gerstenmeier, J. (2000): Die Kluft zwischen Wissen und Handeln. Göttingen.

McDaniel, M.A./Nguyen, N. T. (2001): Situational judgment tests: A review of practice and constructs assessed. In: International Journal of Selection and Assessment, 9, 103-113.

McDaniel, M. A./Whetzel, D. L./Hartman, N.S./Nguyen, N. T./Grubb, W. L. (2014): Situational Judgment Tests: Validity and an Integrative Model. In: Weekley, J.A./Ployhart, R. E. (Hrsg.): Situational Judgment Tests. Theory, Measurement and Application. New York, 183-204.

McDaniel, M.A./Hartmann, N.S./Whetzel, D.L./Grubb, W.L. (2007): Situational Judgment Tests, Response Instructions, and Validity: A Meta-Analysis. In: Personnel Psychology, 60, H. 1, 63-91. https://doi.org/10.1111/j.1744-6570.2007.00065.x.

Moosbrugger, H./Brandt, H. (2020): Itemkonstruktion und Antwortverhalten. In: Moosbrugger, H./Kelava, A. (Hrsg.): Testtheorie und Fragebogenkonstruktion. 3. Aufl. Berlin, 67-90.

Muck, P. M. (2013): Evidenzbasierte Entwicklung von Situational Judgment Tests: Konzeptionelle Überlegungen und empirische Befunde. In: Zeitschrift für Arbeits- und Organisationspsychologie, 57, 185-205.

Mumford, T. V./Van Iddekinge, C. H./Morgeson, F. P./Campion, M.A. (2008): The Team Role Test: Development and Validation of a Team Role Knowledge Situational Judgment Test. In: Journal of Applied Psychology, 93, 250-267. https://doi.org/10.1037/0021-9010.93.2.250.

Nickolaus, R. (2010): Erklärungsmodelle für die Entwicklung der Fachkompetenz – Anmerkungen zu ihren Geltungsansprüchen und didaktischen Implikationen. In: Zeitschrift für Berufs- und Wirtschaftspädagogik, 106, H. 4, 481-490.

O’Connor, B.P. (2000): SPSS and SAS programs for determining the number of components using parallel analysis and Velicer’s MAP test. In: Behavior Research Methods, Instrumentation, and Computers, 32, 396-402. https://doi.org/10.3758/BF03200807.

OECD (2020): PISA 2018 Results (Volume IV): Are Students Smart about Money? PISA, OECD Publishing. Online: https://read.oecd-ilibrary.org/education/pisa-2018-results-volume-iv_48ebd1ba-en#page1 (27.06.2024).

Pospeschill, M. (2022): Testtheorie, Testkonstruktion, Testevaluation. 2. Aufl. München.

Ray, J. J./Najman, J. M. (1986): The Generalizability of Deferment of Gratification. In: The Journal of Social Psychology, 126, H. 1, 117-119. https://doi.org/10.1080/00224545.1986.9713578.

Reinders, H./Post, I. (2022): Testverfahren. In: Reinders, H./Bergs-Winkels, D./Prochnow, A./Post, I. (Hrsg.): Empirische Bildungsforschung. Eine elementare Einführung. Wiesbaden, 175-193.

Remund, D. L. (2010): Financial Literacy explicated: The case for a clearer definition in an increasingly complex economy. In: The Journal of Consumer Affairs, 44, H. 2, 276-295.

Renkl, A. (1996): Träges Wissen. Wenn Erlerntes nicht genutzt wird. In: Psychologische Rundschau, 47, H. 2, 78-92.

Rieger, M.O. (2020): How to Measure Financial Literacy. In: Journal of Risk and Financial Management, 13, H. 12, 1-14. https://doi.org/10.3390/jrfm13120324.

Rost, J. (2004): Lehrbuch Testtheorie – Testkonstruktion. 2. Aufl. Bern.

Schill, M. (2021): Financial Literacy bei jungen Erwachsenen: eine Interviewstudie zur Identifizierung von Gründen, weshalb das monatliche Einkommen nicht ausreicht. Unveröffentlichte Masterarbeit an der Goethe Universität Frankfurt.

Schmeiser, M. D./Seligman, J.S. (2013): Using the Right Yardstick: Assessing Financial Literacy Measures by Way of Financial Well-Being. In: The Journal of Consumer Affairs, 47, H. 2, 243-262. https://doi.org/10.1111/joca.12010.

Schmitt, N./Chan, D. (2014): Situational Judgement Tests: Method or Construct? In: Weekley, J.A./Ployhart, R.E. (Hrsg.): Situational Judgement Tests. Theory, Measurement and Application. New York, 135-155.

Schnell, R./Hill, P.B./Esser, E. (2018): Methoden der empirischen Sozialforschung. Aufl. Berlin/Bosten.

Schuler, A./Brahm, T. (2021): Financial Literacy in den Lehrplänen deutscher Schulen – eine bundeslandübergreifende Analyse. In: Zeitschrift für ökonomische Bildung, H. 10, 1-63. https://doi.org/10.7808/zfoeb.2021.10.77.

Seeber, S./Fischer, A./Michaelis, C./Müller, J. (2014): Zur Messung von Kompetenzen zum nachhaltigen Wirtschaften mit einem Situational Judgement Test. berufsbildung, H. 146, 6-9.

Seeber, S./Nickolaus, R. (2010): Kompetenzmessung in der beruflichen Bildung. In: BWP Berufsbildung in Wissenschaft und Praxis, H. 1, 10-13.

Staub-Kaminsky, I. (2020): Measuring Adolescents‘ Financial Literacy: A Think-Aloud Study. Unveröffentlichte Masterarbeit an der Goethe Universität Frankfurt.

Streiner, D.L. (2003): Starting at the Beginning: An Introduction to Coefficient Alpha and Internal Consistency. In: Journal of Personality Assessment, 80, 99-103. https://doi.org/10.1207/S15327752JPA8001_18.

Überla, K. (1971): Faktorenanalyse. Eine systematische Einführung für Psychologen, Mediziner, Wirtschafts- und Sozialwissenschaftler. Berlin/Heidelberg.

Wagner, E. (2023): Financial Literacy von sächsischen Schülerinnen und Schülern – eine empirische Studie zu ausgewählten Validierungsaspekten des Situational Judgement Tests. Unveröffentlichte Masterarbeit an der Universität Leipzig.

Walstad, W.B./Rebeck, K. (2017): The Test of Financial Literacy: Development and measurement characteristics. In: Journal of Economic Education, 48(2), 113-122. https://doi.org/10.1080/00220485.2017.1285739.

Walstad, W. B./Rebeck, K. (2018): The measurement properties of the Basic Finance Test for children and the Test of Financial Knowledge for youth. In: Empirische Pädagogik, 32, H. 3-4, 248-271.

Weekley, J. A./Ployhart, R. E./Holtz, B. C. (2014): On the Development of Situational Judgement Tests. Issues in Item Development, Scaling and Scoring. In: Weekley, J. A./Ployhart, R. E. (Hrsg.): Situational Judgement Tests. Theory, Measurement and Application. New York, 157-182.

Whetzel, D.L./McDaniel, M.A. (2009): Situational judgment tests: An overview of current research. In: Human Resource Management Review, 19, 188-202.

Whitehead, A.N. (1929): The aims of education and other essays. New York.

Wuttke, E./Aprea, C. (2018): Situational judgment approach for measuring young adults' Financial Literacy. In: Empirische Pädagogik, 3, H. 3-4, 272-292.

Wuttke, E./Seeber, S./Geiser, C./Turhan, L. (2022): Zur Problemhaltigkeit von Aufgaben in kaufmännischen Abschluss- und Zwischenprüfungen – Ergebnisse aus Aufgabenanalysen. In: Zeitschrift für Berufs- und Wirtschaftspädagogik 118, H. 1, 25-52. https://doi.org/10.25162/zbw-2022-0002.

Wuttke, E./Siegfried, C./Aprea, C. (2020): Measuring financial literacy with a Situational Judgement Test: do some groups really perform worse or is it the measuring instrument? In: Empirical Research in Vocational Education and Training, 12, 1-21. https://doi.org/10.1186/s40461-020-00103-x.

Zitieren dieses Beitrags

Wagner, E./Wuttke, E./Happ, R. (2024): Valide Messung von Financial Literacy - Überarbeitung und Analyse eines Situational Judgement Tests. In: *bwp@* Profil 10: Herausforderungen und Gestaltungsfragen für die berufliche Bildung. Digitale Festschrift für Susan Seeber zum 60. Geburtstag, hrsg. v. Michaelis, C./Busse, R./Wuttke, E./Fürstenau, B., 1-24. Online: https://www.bwpat.de/profil10_seeber/wagner_etal_profil10.pdf (24.11.2024).

Die Autor:innen

ELISA WAGNER

Universität Leipzig, Institut für Wirtschaftspädagogik

Grimmaische Str. 12, 04109 Leipzig

ewagner@wifa.uni-leipzig.de

https://www.wifa.uni-leipzig.de/institut-fuer-wirtschaftspaedagogik/team

Prof. Dr. EVELINE WUTTKE

Goethe-Universität Frankfurt, Wirtschaftspädagogik, insb. empirische Lehr-Lern-Forschung

Theodor-W.-Adorno-Platz 1, 60323 Frankfurt

wuttke@em.uni-frankfurt.de

https://www.old.wiwi.uni-frankfurt.de/abteilungen/wipaed/professoren/wuttke/team/prof-dr-eveline-wuttke.html

Prof. Dr. ROLAND HAPP

Universität Leipzig, Institut für Wirtschaftspädagogik

Grimmaische Str. 12, 04109 Leipzig

happ@wifa.uni-leipzig.de

https://www.uni-leipzig.de/personenprofil/mitarbeiter/pd-dr-roland-happ