4.2 Konstruktionskriterien des Testinstruments in der Expertenvalidierung

Nina Johanna Welsandt, Fenna Henicz, Fabio Fortunati, Esther Winther & Hermann Josef Abs

4.2.1 Stichprobe

Expertengestützte Validitätsprüfung ECON 2022

Für die Überprüfung der Validität wurden Konstruktionskriterien des Instrumentariums in eine Expertenbefragung gegeben. Die Expert*innen haben die konstruierten Testitems einerseits entlang schwierigkeitsgenerierender Merkmale bewertet, um ex ante Vorstellungen davon zu entwickeln, welches Leistungsspektrum das Instrumentarium erfassen kann. Andererseits wurden Urteile über spezifische Testcharakteristika – hier: authentische Item- und Testgestaltung sowie Aspekte der Usability – eingeholt. Die Expert*innen (n = 25) repräsentieren Fachwissen aus drei Handlungsfeldern; sie bringen Expertisen aus den Forschungsbereichen Testentwicklung (n = 10), Wirtschaftswissenschaften/Wirtschaftspädagogik beziehungsweise Wirtschaftspsychologie (n = 11) sowie Schule und Unterricht (n = 12) ein.

Multidisziplinäre Expertisen der Expert*innen

4.2.2 Ratings der schwierigkeitsgenerierenden Merkmale

Ein Ex-ante-Rating der Itemschwierigkeiten

Ein Ex-ante-Rating der Itemschwierigkeiten erhöht die Wahrscheinlichkeit, das zu erfassende Konstrukt in angemessener Breite abbilden zu können. Die Expert*innen haben die Itemschwierigkeiten entlang dreier Merkmale bewertet: (1) inhaltliche Spezifität, (2) kognitive Beanspruchung sowie (3) funktionale Modellierung (Klotz et al., 2015; Winther, 2010; Beck, 2020):

- Die inhaltliche Spezifität bewertet Testaufgaben hinsichtlich des zu ihrer Lösung benötigten (Fach-)Wissens. Die Aufgabenkonstruktion wird hier auf *Inhaltsebene* beurteilt.
- Die Art der kognitiven Beanspruchung bewertet Testaufgaben mit Blick auf die zu ihrer Lösung eingeforderten Leistungsfähigkeiten der Schüler*innen. Die Aufgabenkonstruktion wird hier auf kognitiver *Prozessebene* beurteilt.
- Die funktionale Modellierung bewertet Testaufgaben dahingehend, wie anspruchsvoll es ist, die zur Lösung notwendigen Schritte aus der zugrundeliegenden Anforderungssituation zu extrahieren. Die Aufgabenkonstruktion wird hier auf *Kontextebene* beurteilt.

Konstruktionsprozess

Die systematische Konstruktion von Testaufgaben entlang der o. g. schwierigkeitsgenerierenden Merkmale stellt sicher, dass (1) Annahmen über kognitive Theorien in die Testaufgaben einfließen und (2) Testaufgaben formuliert werden, die gut zwischen den zu testenden Personen trennen können, da sie unterschiedliche Fähigkeitsstufen ansprechen.

Die nachfolgende Abbildung 4.2.1 zeigt das Ratingschema, das von den Expert*innen zur Schwierigkeitsprognose genutzt wurde. Die drei schwierigkeitsgenerierenden Merkmale differenzieren zwischen jeweils drei Schwierigkeitsstufen, wobei Stufe 1 für eine geringe Schwierigkeit und Stufe 3 für eine hohe Schwierigkeit der Testaufgabe steht.

Welsandt, N. J., Henicz, F., Fortunati, F., Winther, E. & Abs, H. J. (2024). Konstruktionskriterien des Testinstruments in der Expertenvalidierung. In E. Winther & H. J. Abs (Hrsg.), *ECON 2022. Ökonomische Bildung in Jahrgang 8: Konzeption, Kompetenzen und Einstellungen.* Waxmann (S. 110–116). CC BY 4.0. https://doi.org/10.31244/9783830998389.4_2

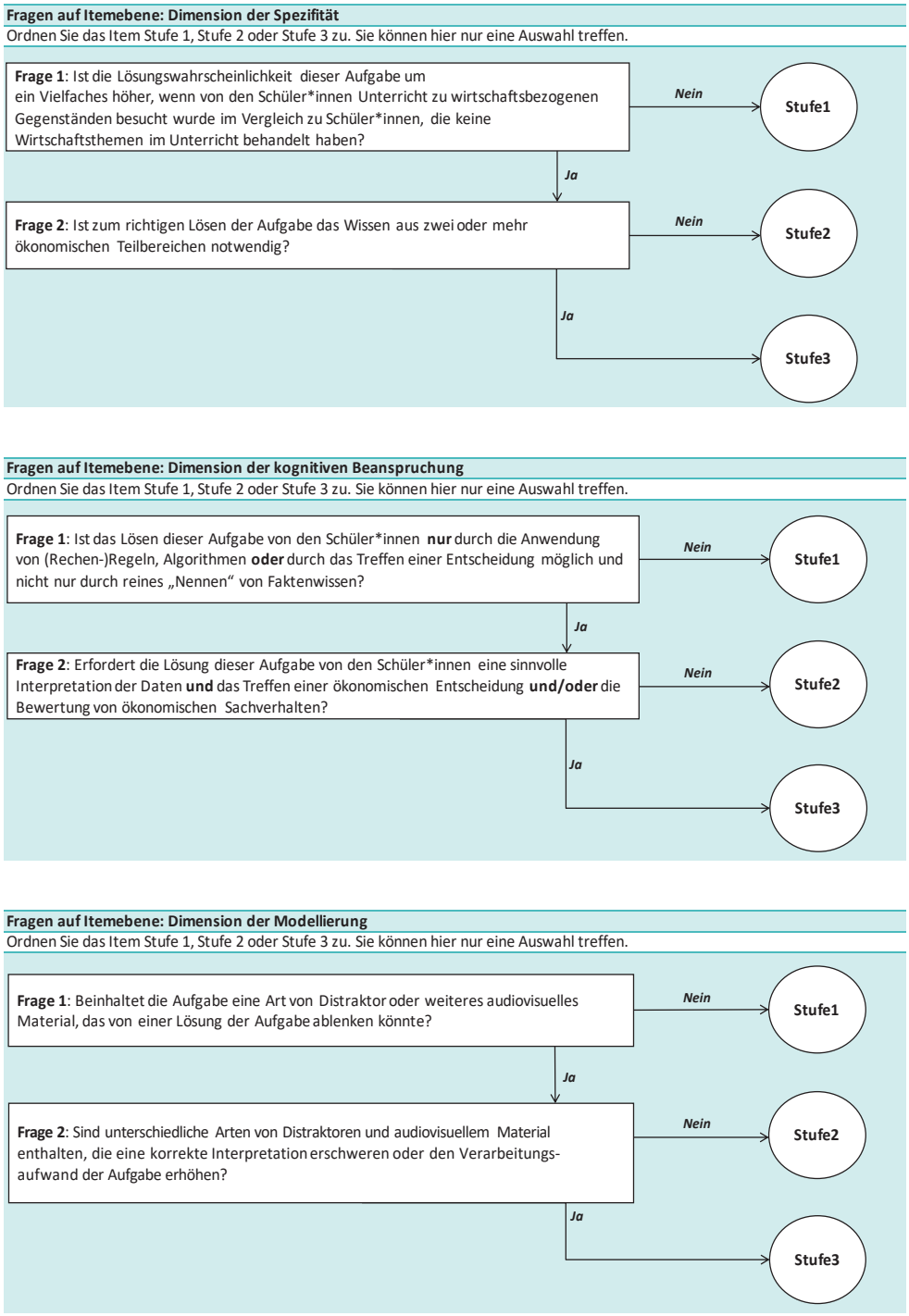


Abbildung 4.2.1: Ratingschema der schwierigkeitsgenerierenden Merkmale

Verteilung der Itemschwierigkeiten

Ein gutes, d.h. zwischen den Fähigkeiten der Schüler*innen hinreichend diskriminierendes, Testinstrument weist eine ausgewogene Verteilung der Itemschwierigkeiten auf. Für den konstruierten Test liegen auf Basis der Urteile der Expert*innen ex ante Schwierigkeitsprognostiken vor, die – wie in Abbildung 4.2.2 dargestellt – ein ausgewogenes Verhältnis von leichten, mittelschweren und schweren Testitems erwarten lassen.

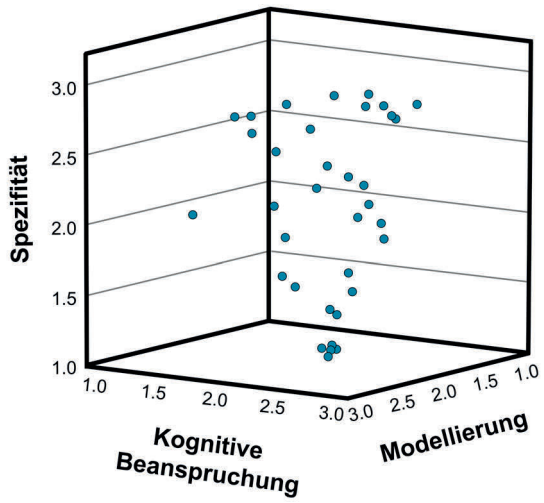
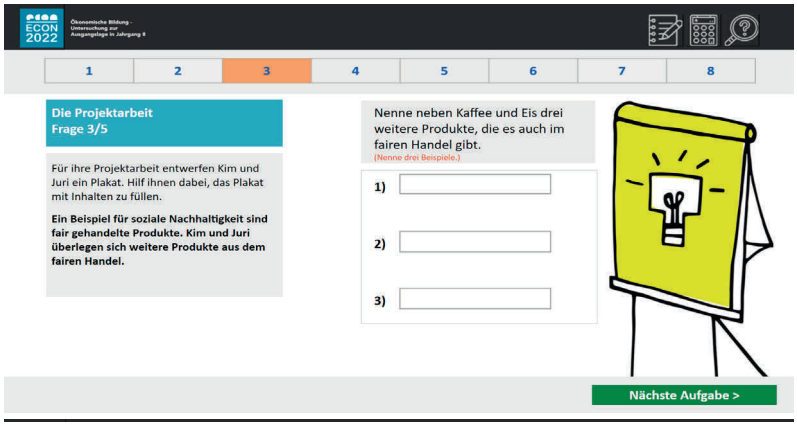


Abbildung 4.2.2: Ex-ante-Beurteilung der Itemschwierigkeiten

Abbildung 4.2.2 illustriert dreidimensional die Bewertungen der schwierigkeitsgenerierenden Merkmale Spezifität, kognitive Beanspruchung und Modellierung. Die durch Punkte dargestellten Werte der x-, y- und z-Achse stellen das durchschnittliche Rating der einzelnen Items dar. Zur Veranschaulichung werden in Abbildung 4.2.3 Beispiele verschieden schwieriger Testitems präsentiert.

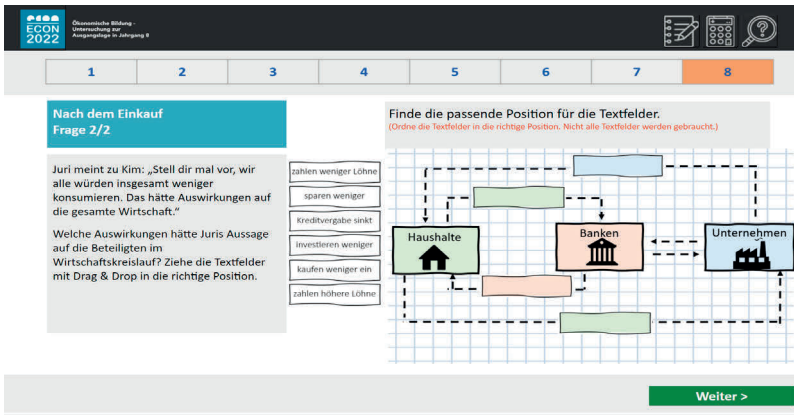
4.2.3 Ratings testcharakteristischer Merkmale

Neben einer schwierigkeitsbeschreibenden Beurteilung wurden die Expert*innen gebeten, testcharakteristische Merkmale zu bewerten. Hierzu war zunächst die authentische Gestaltung der Testinhalte und dann die Usability des Tests einzuschätzen. Zur Bewertung der Authentizität wurde ebenfalls ein dreistufiges Ratingschema genutzt (Abbildung 4.2.4): Es wird davon ausgegangen, dass lebensweltnahe Handlungssituationen eine Aufgabe besser zugänglich machen. Dafür müssen authentische Situationen modelliert werden. Stufe 1 wird erreicht, wenn das Item der Zielgruppe aus dem Alltag vertraut ist. Ist die Situation zumindest theoretisch zugänglich, kann sie Stufe 2 zugeordnet werden. Ein Item fällt unter Stufe 3, wenn nicht erwartet werden kann, dass die Zielgruppe einer solchen Situation im Alltag perspektivisch begegnen kann.



Prognostisch leichtes Testitem 3_3: Fair-Trade Produkte
Die Schüler*innen müssen in einem offenen Format Fair-Trade-Produkte aufzählen. Das Nennen der Produkte, sprich die Anwendung von Faktenwissen, ermöglicht eine korrekte Lösung der Aufgabe. Dabei beinhaltet die Aufgabe keine weiteren Distraktoren. Spezifität (M=1.37), kognitive Beanspruchung (M=1.12) und auch Modellierung (M=1.06) wurden in der durchschnittlichen Bewertung der ersten Stufe zugeordnet und somit als leicht eingestuft.

Prognostisch mittelschweres Testitem 1_2: Bedürfnisse und Bedarf
In dieser Hotspot-Aufgabe können durch Anklicken Produkte von der Einkaufsliste gestrichen werden, die nicht zum täglichen Bedarf zählen. Für das Lösen der Aufgabe ist es hilfreich, Unterricht zu wirtschaftsbezogenen Gegenständen besucht zu haben. Weiterhin reicht das reine Faktenwissen bei dieser Aufgabe nicht mehr aus. Das Testitem wurde in allen Merkmalen auf Stufe 2 geratet: Spezifität (M=1.83), kognitive Beanspruchung (1.73), Modellierung (1.90).



Prognostisch schweres Testitem 8_2: Wirtschaftskreislauf
Über Drag-&-Drop-Felder müssen die Verbindungen im Wirtschaftskreislauf passend hergestellt werden. Dafür muss das Wissen mehrerer Teilbereiche kombiniert angewendet werden. Die bereits vorgegebenen Begriffe sind zu interpretierten und Entscheidungen sind zu treffen. Unterschiedliche Arten von Distraktoren erhöhen die Schwierigkeit zusätzlich. Spezifität (M=2.96), kognitive Beanspruchung (M=2.82) und Modellierung (M=2.48) wurden im Mittel der Stufe 3 zugeordnet.

Abbildung 4.2.3: Beispiele unterschiedlich schwerer Testitems

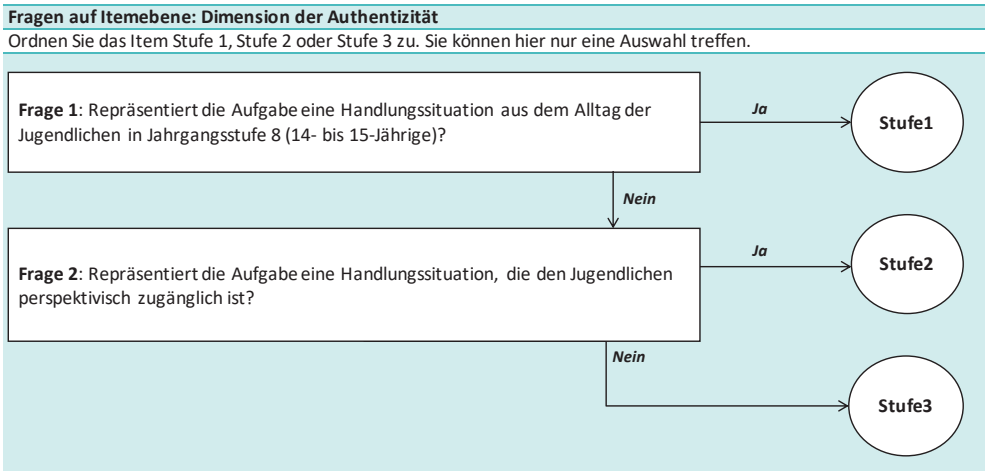


Abbildung 4.2.4: Beurteilung der Authentizität

Das Expertenrating zeigt, dass die Expert*innen die Aufgaben größtenteils als authentisch einschätzten. Eine Ausnahme bilden die Aufgaben 8_1 und 8_2. Die Grenzwerte wurden aus den Bewertungen der Expert*innen abgeleitet. So bilden nur die Aufgaben 8_1 und 8_2 laut den Expert*innen keine perspektivisch zugängliche Handlungssituation für Schüler*innen ab. Die Werte lagen hier über 2.30. 16 Aufgaben erhielten mit Werten zwischen 1.51 und 2.30 eine Bewertung mittlerer Authentizität. 17 Aufgaben wurden mit Werten unter 1.5 als sehr authentisch eingestuft.

Beurteilung der Usability

Zur Beurteilung der Usability auf Testebene wurden den Expert*innen am Ende der Befragung 15 Fragen gestellt (siehe Tabelle 4.2.1), bei denen zu jeder Aussage angegeben werden sollte, inwieweit diese in Bezug auf die Zielgruppe von Achtklässler*innen persönliche Zustimmung findet. Dabei wurde ein vierstufiges Antwortformat mit den Antwortmöglichkeiten 1 = stimme nicht zu, 2 = stimme eher nicht zu, 3 = stimme eher zu und 4 = stimme zu gewählt. Die Expert*innen bewerteten verschiedene Aspekte der Benutzerfreundlichkeit des Assessments.

Die Ergebnisse dieser Expertenbefragung lieferten wertvolle Einblicke in die Usability des TBA-EL aus der Perspektive von Fachexpert*innen. Die meisten Aspekte der Benutzerfreundlichkeit wurden positiv bewertet, während es einige Bereiche gibt, die weiterhin Aufmerksamkeit erfordern. Diese Ergebnisse bieten eine Grundlage für gezielte Verbesserungen des Assessments und eine Optimierung seiner Benutzerfreundlichkeit. Die Ergebnisse dieser Befragung werden im Folgenden präsentiert:

Befunde zu den Bewertungen der Usability

- Intuition des Userinterfaces: Die Auswertung von 21 Bewertungen ergab einen Durchschnittswert (M) von 3.71. Dies deutet darauf hin, dass die Expert*innen die Nutzung des Userinterfaces im Allgemeinen als intuitiv empfinden.
- Intuition der Funktionen einzelner Buttons: Auch hier wurde ein Durchschnittswert von 3.71 ermittelt. Dies legt nahe, dass die Expert*innen die Funktionen der Buttons überwiegend als intuitiv wahrnehmen.
- Verständnis eingesetzter Gestaltungselemente: Hier ergab sich basierend auf den Einschätzungen von 21 Expert*innen ein Durchschnittswert von 3.57, der darauf schließen lässt, dass die Gestaltungselemente für die Zielgruppe angemessen ausgewählt wurden.
- Verständnis der zu bearbeitenden Aufgaben: Basierend auf den Einschätzungen von 21 Expert*innen wurde hierbei ein Durchschnittswert von 3.62 er-

mittelt. Die zu bearbeitenden Aufgaben wurden demnach zielgruppengerecht eingeführt und erstellt.

- Häufigkeit von Überraschungen während der Aufgabenbearbeitung: Hier zeigte sich basierend auf den Einschätzungen von 20 Expert*innen ein Durchschnittswert von 2.05.
- Mehrwert des Erklärvideos als Unterstützung: Die Auswertung von 20 Bewertungen ergab einen Durchschnittswert von 3.15. So wurde der Aussage „Durch das Erklärvideo fällt mir der Umgang mit dem Assessment leichter" durchschnittlich nur eher zugestimmt. Infolgedessen wurde das Erklärvideo noch mal auf Funktionalität überprüft und angepasst.
- Offensichtlichkeit relevanter Informationen für die Aufgabenbearbeitung: Hierbei wurde basierend auf den Einschätzungen von 20 Expert*innen ein Durchschnittswert von 3.35 ermittelt. So konnte aus der Befragung abgeleitet werden, dass die relevanten Informationen teilweise nicht direkt ersichtlich gewesen sind. Dies führte zu einer Überprüfung und Anpassung des Aufgabendesigns.
- Ablenkung durch Gestaltungselemente. Die Auswertung von 21 Bewertungen ergab einen Durchschnittswert von 3.0. Gestaltungselemente wurden demnach teilweise als ablenkend empfunden und wurden daher überarbeitet. In Anbetracht der Modellierung einer Aufgabe dürfen Gestaltungselemente nur als Ablenkung empfunden werden, wenn diese gezielt als Distraktoren eingesetzt wurden. Ist dies nicht der Fall, müssen die Gestaltungselemente verändert werden.
- Lesbarkeit der Arbeitsaufträge des Assessments sowie der verwendeten Materialien: Die Auswertung ergab basierend auf den Einschätzungen von 21 Expert*innen Durchschnittswerte von 3.57 und 3.19. Aus dem Durchschnittswert lässt sich ableiten, dass Elemente nicht durchweg gut lesbar waren und dass eingesetzte Materialien unübersichtlich waren. Auch hier musste es Anpassungen geben.
- Lesbarkeit der Untertitel in den Videoelementen des Assessments: Hierbei wurde basierend auf den Einschätzungen von 20 Expert*innen ein Durchschnittswert von 3.55 ermittelt.
- Übersichtlichkeit der eingesetzten Materialien: Dies ergab basierend auf den Einschätzungen von 21 Expert*innen einen Durchschnittswert von 3.38.
- Erleichterung der Aufgaben durch gewähltes Design: Hierbei wurde basierend auf den Einschätzungen von 20 Expert*innen ein Durchschnittswert von 3.2 ermittelt. Das Design kann die Bearbeitung einer Aufgabe inhaltlich nicht vereinfachen. Trotzdem sollte es eine Aufgabe auch nicht erschweren.
- Verständlichkeit für Schüler*innen der 8. Klasse: Dies ergab basierend auf den Einschätzungen von 21 Expert*innen einen Durchschnittswert von 3.38.
- Angemessenheit der Distraktoren für Schüler*innen der 8. Klasse: Die Auswertung ergab basierend auf den Einschätzungen von 21 Expert*innen einen Durchschnittswert von 3.19. Die Einschätzung der Expert*innen wies darauf hin, dass Schüler*innen nicht vollkommen klar sein könnte, was bei den Aufgaben von ihnen gefordert wird. Dies geht mit der Aussage, dass die Distraktoren für Achtklässler*innen nur eher als angemessen empfunden wurden, einher. Auch hier wurde überprüft, welche der Distraktoren angepasst werden können.

Tabelle 4.2.1: Gesamtbewertung der Usability-Fragen auf Testebene

	Expertenrating (ER)				
	1 = stimme nicht zu, 2 = stimme eher nicht zu, 3 = stimme eher zu, 4 = stimme zu				
	N	Min	Max	M	Std. Abweichung
Die Nutzung des Userinterfaces ist intuitiv.	21	3	4	3.71	0.46
Die Funktionen der einzelnen Buttons sind intuitiv.	21	3	4	3.71	0.46
Die eingesetzten Gestaltungselemente habe ich schnell verstanden.	21	2	4	3.57	0.60
Wie ich eine Aufgabe bearbeiten soll, habe ich schnell verstanden.	21	2	4	3.62	0.59
Bei der Bearbeitung der Aufgaben war ich oft überrascht.	20	1	4	2.05	1.19
Durch das Erklärvideo fällt mir der Umgang mit dem Assessment leichter.	20	1	4	3.15	0.99
Für die Bearbeitung der Aufgaben sind alle relevanten Informationen ersichtlich.	20	2	4	3.35	0.75
Die eingesetzten Gestaltungselemente haben mich bei der Bearbeitung der Aufgaben nicht abgelenkt.	21	1	4	3.00	0.89
Die Arbeitsaufträge des Assessments sind gut lesbar.	21	2	4	3.57	0.60
Die Materialien sind gut lesbar.	21	1	4	3.19	0.87
Die Untertitel der Videoelemente sind gut lesbar.	20	1	4	3.55	0.83
Die eingesetzten Materialien im Assessment sind übersichtlich gestaltet.	21	2	4	3.38	0.59
Das Design der Aufgaben hat mir das Bearbeiten erleichtert.	20	2	4	3.20	0.62
SuS der 8. Klasse ist klar, was von ihnen in der Aufgabe gefordert wird.	21	2	4	3.38	0.67
Die Distraktoren sind für SuS der 8. Klasse angemessen.	21	2	4	3.19	0.75

4.2.4 Implikationen der Ergebnisse

Die Datenanalyse der Expertenbefragung zeigt, dass sie sich trotz unterschiedlicher Expertise in ihrer Bewertung der Spezifität, kognitiven Beanspruchung, Authentizität und Modellierung weitgehend einig sind. Basierend auf den Ergebnissen der Expertenbefragung und der Validierung der Feldtestdaten (vgl. Kapitel 4.3) wurde das Prüfinstrument überarbeitet. Besonders schwierige Items wurden zu leichteren Items abgeändert. Distraktoren wurden bspw. der Altersgruppe entsprechend angepasst. Weitergehend wurden überflüssige Elemente bei der Überarbeitung der Modellierung des Assessments entfernt.

In Bezug auf die negativen Authentizitätseinschätzungen des Expertenratings wurden die Aufgaben 8_1 und 8_2 nochmal genauer überprüft. Da es sich bei den beiden Aufgaben um die letzten des Assessments handelte, wurde hier entschieden, den Lebensweltbezug zu lockern, um Aufgaben zu generieren, die stärker auf die Reflexion über wirtschaftliche Systeme ausgerichtet waren.

Literatur

Beck, K. (2020). Ensuring content validity of psychological and educational tests – the role of experts. *FLR*, 1–37. https://doi.org/10.14786/flr.v8i6.517

Klotz, V. K., Winther, E. & Festner, D. (2015). Modeling the development of vocational competence: A psychometric model for economic domains. *Vocations and Learning*, *8*(3), 247–268. https://doi.org/10.1007/s12186-015-9139-y.

Winther, E. (2010). *Kompetenzmessung in der beruflichen Bildung.* Habilitation. Bertelsmann.