

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Klärung des alltagsweltlichen Konzeptes der
Vertrauenswürdigkeit (von Messdaten) bei Lernenden der
Sekundarstufe 1 und Prüfung der Langzeitwirksamkeit von
darauf basierenden Unterrichtskonzeptionen

verfasst von | submitted by
Eva Abl BEd

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of
Master of Education (MEd)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree
programme code as it appears on the
student record sheet:

UA 199 520 523 02

Studienrichtung lt. Studienblatt | Degree
programme as it appears on the student
record sheet:

Masterstudium Lehramt Sek (AB) Unterrichtsfach
Mathematik Unterrichtsfach Physik

Betreut von | Supervisor:

Univ.-Prof. Dr. Martin Hopf

Mitbetreut von | Co-Supervisor:

Mag. Dr. Clemens Nagel

Eidesstattliche Erklärung

Ich erkläre hiermit an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Böheimkirchen, am 23. August 2024

Eva Abl, BEd

Ein großes Dankeschön an...

...meinen Betreuer *Clemens Nagel*, der mich über die ganze Zeit hinweg auf die beste Art und Weise betreut hat.

...das AECC für alle guten Inputs und Verbesserungsvorschläge.

...alle Schüler*innen, die sich bereit erklärt haben an Brainstormings, Interviews und dem Ausfüllen von Fragebögen teilzunehmen. Auch danke an alle Kolleg*innen, die dies erst möglich gemacht haben.

...Betül, Nicolas, Raphaela und Thomas, die mich bei sämtlicher Forschung tatkräftig unterstützt haben, Formulierungen ausgearbeitet haben, Fragebögen von A nach B getragen haben, Kreuzerl abgetippt haben, etc.

...Katt, die sich sämtliche Sorgen und Beschwerden bei jedem Fladerei-Essen angehört hat.

...Anna, mein persönliches Wörterbuch, die immer Wörter, egal ob Deutsch oder Englisch, für mich parat hatte, wenn sie mir nicht eingefallen sind.

...Lena, die Stunden mit mir an einem Tisch verbracht hat, damit ich diese Arbeit fertigstellen konnte. Danke, ohne dich wäre dieses Studium so nie möglich gewesen.

Abstract

Die vorliegende Masterarbeit untersucht die nachhaltigen Auswirkungen zweier unterschiedlicher Unterrichtskonzeptionen zu Messunsicherheiten auf das Verständnis von Schülerinnen der Sekundarstufen I und II. Zudem wird der Begriff „Vertrauenswürdigkeit“ im Kontext der Schülerinnen der Sekundarstufe I beleuchtet. Die Methodik umfasste einen zeitverzögerten Test (Follow-Up-Test) zur Messung langfristiger Lerneffekte und die Entwicklung eines Fragebogens zur Erfassung des Verständnisses von Vertrauenswürdigkeit.

Der Follow-Up-Test, der 1-2 Jahre nach der Unterrichtsintervention durchgeführt wurde, ergab, dass die Kernkonzepte der Messunsicherheit bei den meisten Schülerinnen weiterhin präsent sind. Die Analyse der Testergebnisse von 41 Schülerinnen zeigte, dass die Gruppen, die die Unterrichtskonzeptionen von Hannah Loidl und Rupert Bärenthaler-Pachner durchlaufen hatten, keine signifikanten Leistungsunterschiede aufwiesen. Dies deutet darauf hin, dass beide Konzepte effektiv sind, obwohl Unterschiede im Fokus und Alter der Schüler*innen die Ergebnisse beeinflusst haben könnten.

Die Entwicklung des Fragebogens erfolgte in zehn Phasen und beinhaltete Experteninterviews, Brainstormings und Pretests. Die endgültige Version des Fragebogens wurde hinsichtlich ihrer Reliabilität überprüft. Die Ergebnisse zeigen, dass Schüler*innen der Sekundarstufe I ein grundlegendes Verständnis für den Begriff der Vertrauenswürdigkeit besitzen. Sie bewerteten Quellen gemäß ihrer Alltagserfahrungen, wobei technische Hilfsmittel wie Taschenrechner als vertrauenswürdiger eingeschätzt wurden als unbekannte Quellen. Dies unterstreicht die Bedeutung des Brückenkonzepts der Vertrauenswürdigkeit für den Unterricht.

Die Arbeit zeigt, dass beide Unterrichtskonzeptionen zu nachhaltigen Lerneffekten führen und dass das Verständnis von Vertrauenswürdigkeit bei den Schüler*innen bereits gut entwickelt ist. Diese Erkenntnisse bieten wertvolle Impulse für die weitere Gestaltung von Unterrichtsmaterialien zur Messunsicherheit und zur Förderung kritischen Denkens.

Inhaltsverzeichnis

1	Einleitung	6
1.1	Hintergrund und Motivation	6
1.2	Ziel und Forschungsfragen	6
2	Theoretischer Hintergrund	8
2.1	Vertrauenswürdigkeit als gutes Brückenkonzept für Messunsicherheiten . .	8
2.1.1	Messunsicherheit nach der Methode Typ A	8
2.1.2	Messunsicherheit nach der Methode Typ B	8
2.1.3	Vertrauenswürdigkeit	8
2.2	Die Four-Building-Blocks-Methode	10
2.2.1	Construct Map	10
2.2.2	Item Desgin	11
2.2.3	Outcome Space	11
2.2.4	Measurement Model	11
3	Methodik - Zeitverzögerter Test	12
3.1	Erstellen des Tests	12
3.2	Durchführung des Tests	19
4	Methodik - Fragebogen zur Vertrauenswürdigkeit	19
4.1	Entwicklung des Fragebogens	19
4.1.1	Aufsuchen der Dimensionen und erster Fragebogenentwurf	20
4.1.2	Experteninterview	21
4.1.3	Experten-Diskussion	22
4.1.4	Pretesting 1 - Brainstorming in sechs Klassen	22
4.1.5	Überarbeitung in Gruppendiskussion - Reduktion der Dimensionen	23
4.1.6	Pretesting 2 - Think Aloud Interviews	26
4.1.7	Finaler Fragebogen	32
4.2	Qualitätssicherung	35
4.2.1	Intercoder Reliabilität Testung	35
4.2.2	Stichprobenwahl und Testing-Phase	38
4.2.3	Die Faktorenanalyse - Validität	38
4.2.4	Cronbachs Alpha - Reliabilität	39
4.2.5	Objektivität	40
4.2.6	Weitere Gütekriterien	40
5	Ergebnisse	42
5.1	Ergebnisse des zeitverzögerten Tests	42
5.2	Ergebnisse des Fragebogens zur Vertrauenswürdigkeit	46

5.2.1	Synonym- und Gegenteilabfrage	46
5.2.2	Analyse der Konstrukte	47
5.2.3	Vertrauenswürdigkeit von Quellen	52
5.2.4	Vertrauenswürdigkeit von Messungen	52
6	Diskussion	54
6.1	Zeitverzögerter Test	54
6.1.1	Limitationen des Tests	55
6.2	Fragebogen zur Vertrauenswürdigkeit	56
6.2.1	Limitationen des Fragebogens	57
7	Literatur	58
A	Erster Fragebogenentwurf	61
B	Überarbeitung: Erster Fragebogenentwurf	65
C	Zweiter Fragebogenentwurf	68
D	Informationen an die Erziehungsberechtigten	70
E	Einverständniserklärung	71

1 Einleitung

1.1 Hintergrund und Motivation

Messunsicherheiten sind ein wichtiges Thema der Physik, da eines der wichtigsten Werkzeuge der Erkenntnisgewinnung in der Physik die empirische Forschung bildet. Anders gesagt ist die Messung ein zentraler Gegenstand dieser Naturwissenschaft. Mit Messungen kommen auch Messunsicherheiten einher. Der Begriff der Messunsicherheit kann Schüler:innen Probleme bereiten, da der Begriff „unsicher“ mit anderen Begriffen wie „Fehler“ oder „falsch“ in Verbindung gebracht wird (Loidl, 2021).

An dieser Stelle haben bereits Hannah Loidl und Rupert Bärenthaler-Pachner mit ihren Masterarbeiten angesetzt. Sie haben Unterrichtskonzeptionen für die Sekundarstufen I und II erstellt, in denen sie den Begriff *Messunsicherheit* über den Begriff der *Vertrauenswürdigkeit* einführen. Diese Unterrichtskonzeptionen mit dem zentralen domänenspezifischen Designprinzip der Vertrauenswürdigkeit als Brückenkonzept zwischen Alltag und Fachkonzept der Messunsicherheiten zeigten damals ein erfolgreiches Erreichen der gesetzten Lernziele. In dieser Arbeit sollen nun zwei Aspekte überprüft werden. Einerseits soll die Nachhaltigkeit der beiden Unterrichtskonzeptionen, mittels einem schriftlichen Test, analysiert werden und andererseits soll der Begriff der *Vertrauenswürdigkeit* selbst von mehreren Seiten beleuchtet werden, um so feststellen zu können, in wie fern er als Brückenkonzept zur *Messunsicherheit* geeignet ist.

1.2 Ziel und Forschungsfragen

Aus dem thematischen Kontext ergeben sich folgende Fragen:

Hauptfragen:

1. Wirken sich die beiden bestimmten Unterrichtskonzeptionen zu **Messunsicherheiten** nachhaltig (1-2 Jahre) auf das Verständnis der Schüler*innen der SEK I und SEK II zu *Messunsicherheiten* aus?
2. Was bedeutet der Begriff *Vertrauenswürdigkeit* für Schüler*innen der Sekundarstufe I?

Teilfragen:

- 1.1 Können Schüler*innen der SEK I und SEK II nach 1-2 Jahren die entsprechenden Key-Ideas im gegebenen Kontext anwenden?

- 2.1 Welches durch alltagsweltliche Präkonzepte indizierte gedankliche Konstrukt steckt hinter dem Begriff *Vertrauenswürdigkeit*?
- 2.2 Ist das Begriffskonstrukt in geeigneter Weise anknüpfungsfähig für den Vertrauenswürdigkeitsbegriff von Messungen?
- 2.3 Wie gut lässt sich das Konstrukt mit dem Begriff Messunsicherheit verbinden?
- 2.4 Werden Medien/Personen/Institutionen/... die Eigenschaften *vertrauenswürdig* oder *nicht vertrauenswürdig* von Schüler*innen der SEK I zugesprochen?
- 2.5 (In Abhängigkeit von der vorherigen Frage) Welche Eigenschaften müssen Medien/Personen/Institutionen besitzen, damit sie von Schüler*innen der SEK I als *vertrauenswürdig* eingestuft werden?
- 2.6 Wie kann der Begriff der *Vertrauenswürdigkeit* in der Physikunterricht, im Zuge der Messunsicherheiten, eingebracht werden?

2 Theoretischer Hintergrund

2.1 Vertrauenswürdigkeit als gutes Brückenkonzept für Messunsicherheiten

2.1.1 Messunsicherheit nach der Methode Typ A¹

Die Typ-A-Messunsicherheit ist die Unsicherheit, die bei der Auswertung einer Messreihe mittels statistischer Verfahren ermittelt wird. Daraus folgt, dass einzelne Messwerte auch keine solche Messunsicherheiten besitzen. Die Typ-A-Messunsicherheiten hängen nicht mit anderen Messunsicherheiten zusammen, sie sind also *unkorreliert*. In den Unterrichtskonzeptionen von Loidl und Bärenthaler-Pachner werden die Typ-A-Messunsicherheiten mit dem arithmetischem Mittel und der Standardabweichung berechnet, aber auch die Spannweite wird zur Abschätzung der Vertrauenswürdigkeit zugezogen (vgl. Nagel, 2021).

2.1.2 Messunsicherheit nach der Methode Typ B²

Im Gegensatz zur Typ-A-Messunsicherheit wird diese Messunsicherheit bereits bei einzelnen Messwerten bestimmt und kann mit anderen Messunsicherheiten korrelieren. Typ-B-Messunsicherheiten sind unter anderem solche, die sich auf das Messgerät zurückführen lassen. Es werden drei Haupttypen unterschieden:

Die *Eichunsicherheit* ergibt sich durch die Kalibrierung des Messgeräts an einem internationalen Standard. Die *Linearitätsunsicherheit* eines vorab kalibrierten Messgeräts gibt Aufschluss darüber, wie genau die Skalenteile des Instruments die gemessenen Werte widerspiegeln. Aufgrund potenzieller Unterschiede in den Unsicherheiten zwischen den einzelnen Skalenteilen ist diese Unsicherheit unkorreliert. Die *Digitalisierungsunsicherheit* (oder Skalenunsicherheit bei analogen Geräten) ist durch die Auflösung von Messgeräten gegeben. Die Skala eines Messgeräts kann nur eine endliche Anzahl von Unterteilungen haben, jedoch muss der wahre Wert nicht genau auf einer Unterteilung liegen, sondern kann im gesamten Bereich der Skala liegen (Nagel, 2021).

2.1.3 Vertrauenswürdigkeit

Hopf, Schecker, Höttecke und Wiesner (2022) schreiben in ihrem Grundlagenlehrbuch „Physikdidaktik kompakt“ über das Lernen aus konstruktivistischer Sicht folgende Zusammenfassung: Über unsere Sinneskanäle nehmen wir Sinnesdaten auf, z. B. Schriftzeichen, Tafelbilder, Gestalten, experimentelle Anordnungen, Laute. Diese Sinneseindrücke sind zunächst nichts weiter als bloße Daten, die noch keine Bedeutung für den Empfänger haben. Bedeutung können diese Sinnesdaten nur mit Hilfe der bei den Empfängern bereits

¹Hierbei handelt es sich um die fachlich korrekte Bezeichnung der Methode, welche aber in weiterer Folge Typ-A-Messunsicherheit genannt wird.

²Auch hier handelt es sich um die fachlich korrekte Bezeichnung der Methode, welche aber in weiterer Folge Typ-B-Messunsicherheit genannt wird.

vorhandenen Denk- und Wissensstrukturen erhalten, die im Langzeitgedächtnis gespeichert sind:

„Der Zuhörer muss dem, was er hört, und dem, was er sieht, einen Sinn geben. Und das ist nur möglich, indem er Verbindungen mit schon Bekanntem herstellt, also konstruiert. [...] In einem gewissen Sinn wird die Information vom Aufnehmenden konstruiert, indem er schon gespeicherte Vorstellungen aktiviert, in die er die Eingabe einordnet.“ (Jung, 1986).

In der konstruktivistischen Lerntheorie wird also ein Konzept von Kognition und Lernen vorausgesetzt, das neurophysiologisch ein „geschlossenes Gehirn“, also ein kognitiv abgeschlossenes System, postuliert. Dieses setzt Reize von außen in neuronale Erregung um und interpretiert selbstständig in Abhängigkeit von Kontext und Vorerfahrung (gespeicherte Strukturen, sogenannte Gedächtnisinhalte). Bedeutung kann demnach nicht durch externe Signale in das kognitive System übertragen werden. Bedeutungskonstruktion findet individuell und kontextabhängig immer wieder neu im Gehirn statt. Gedächtnisinhalte stellen „Strukturen“ dar, mit Hilfe derer Bedeutungen erzeugt werden können. Solche Gedächtnisinhalte entstehen durch Bedeutungsentwicklungen. Hierbei müssen bei einer konvergierenden Folge von Bedeutungskonstruktionen die Diskrepanzen zwischen den auf Gedächtnisinhalten basierenden Erwartungen und den neuerlichen individuellen Wahrnehmungen des gleichen Inhalts kleiner werden. Von Glasersfeld (1997) bezeichnet solche, sich bewährenden Prozesse als „viabel“. Es braucht also ein Konzept, das bereits im Alltag der Schülerinnen und Schüler verankert und gut genug ausgeprägt ist, an welches angeknüpft werden kann. So können im Unterricht für die Einführung von Messunsicherheiten viable Lernprozesse als konvergierende Folge von Bedeutungskonstruktionen in Anknüpfung an und ausgehend von einem das Alltagskonzept hin zum Fachkonzept der Messunsicherheiten gestaltet werden.

Rund um das Konzept eines Messwertes, der möglichst nahe dem (unbekannten) wahren Wert liegen soll, existieren zahlreiche ähnliche Begriffe und Konzepte. Im Folgenden soll die Bedeutung, Herkunft und Abgrenzung der *Vertrauenswürdigkeit* aufbereitet werden, um ihre Eignung als Brückenkonzept und Anknüpfungspunkt aus dem Alltag der Schülerinnen und Schüler für die Einführung des Konzepts der Messunsicherheiten zu diskutieren und zu beurteilen.

In der Soziologie definiert Georg Simmel schon vor über 100 Jahren die soziale Form des Vertrauens zwischen Menschen als einen mittleren Zustand zwischen Wissen und Nichtwissen, also um eine „Hypothese künftigen Verhaltens“ (vgl. Endreß 2015). „Vertrauen ist der Wille, sich verletzlich zu zeigen. (Osterloh & Weibl 2006)“ Dieser einfache Satz umfasst mehrere Vertrauensdimensionen: Vertrauen entsteht in Situationen, in denen der

Vertrauende (der Vertrauensgeber) mehr verlieren als gewinnen kann – er riskiert einen Schaden bzw. eine Verletzung. Vertrauen manifestiert sich in Handlungen, die die eigene Verletzlichkeit erhöhen. Und schließlich liefert man sich dem Vertrauensnehmer aus und setzt zum Vertrauenssprung an. Der Grund, warum man sich ausliefert, ist die positive Erwartung, dass der Vertrauensnehmer die Situation nicht zum Schaden des Vertrauensgebers verwendet. Außerdem kann man Vertrauen vergrößern, indem man Informationen gibt oder gewinnt (vgl. Endreß 2015). Das Gegenteil des Vertrauens ist das Misstrauen; es beinhaltet wesentlich, dass man gegenüber anderen Personen, weil man sie negativ bewertet, Vorsichtsmaßnahmen ergreift, um Schädigung durch sie auszuschließen. All diese Dimensionen beziehen sich natürlich auf Vertrauen zwischen Menschen. Wenn man aber über die Eigenschaften der Messunsicherheit als Geber des Vertrauensbereiches, in dem sich mit gewisser Wahrscheinlichkeit der wahre Wert befinden soll, nachdenkt, kann man sie ohne weiteres auch auf die Dimensionen und Eigenschaften der Messunsicherheit umlegen. In Wikipedia ist zu finden: „Vertrauen ist ein Phänomen, das in unsicheren Situationen oder bei risikohaftem Ausgang einer Handlung auftritt: Wer sich einer Sache sicher sein kann, muss nicht vertrauen.“ Nachdem man auf der Suche nach dem wahren Wert von Messungen selbigen niemals sicher finden kann (vgl. Drosig, 2006), weil dies letztlich der Erkenntnistheorie nach Popper (1934) widersprechen würde, muss man zwangsläufig Messergebnissen vertrauen – wie sehr, das hängt vom Grad der gegebenen Informationen über den Vertrauensbereich (die Messunsicherheit) ab.

2.2 Die Four-Building-Blocks-Methode

Die *Four-Building-Blocks*-Methode stellt eine Art Gerüst dar, mit dessen Hilfe es ermöglicht werden soll, ein Messinstrument zu erstellen, indem verstanden wird, welche Konstrukte dahinter liegen. Dabei werden *four building blocks* angewendet, welche im Folgenden genauer beschrieben werden. Der Begriff des Messinstruments ist als eine Technik zu verstehen, die etwas Messbares mit latenten, also nicht direkt messbaren, aber in der Theorie vorhandenen Größen in Verbindung bringt. Die *Four-Building-Blocks*-Methode soll der Leitfaden für solch ein Messinstrument sein. Im Falle dieser Arbeit handelt es sich um einen Fragebogen zum Thema Vertrauenswürdigkeit (Wilson, 2005).

2.2.1 Construct Map

„Eine *Construct Map* stellt die Visualisierung eines bestimmten Konstrukts dar“ (Korner, 2014). Um diese Visualisierung überhaupt umsetzen zu können, muss sich zuerst überlegt werden, wie man die unterschiedlichen Aspekte überhaupt evidenzbasiert hinterfragen kann. Die eigentliche *Construct Map* wird nun mithilfe der unterschiedlichen Charakteristika eines Aspekt erstellt. Hierbei wird vorausgesetzt, dass sich diese Charakteristika überhaupt ordnen lassen in eine Reihenfolge à la „sehr stark positiv - neutral - sehr stark negativ“ und

sich im weiteren Sinne fast stetig von dem einen Extremum zum anderen zerlegen lassen (vgl. Korner, 2014). Die große Schwierigkeit hier wird sein, Charakteristika zu formulieren, die sich tatsächlich evidenzbasiert abbilden lassen.

2.2.2 Item Desgin

Auf Basis der *Construct Map* können nun Items erstellt werden, die dann im entgültigen Evaluationsbogen enthalten sind. Dies passiert im zweiten Schritt des *Item Designs*. Die Wahl der Items geht einher mit der Wahl der Skala, die im endgültigen Fragebogen verwendet wird. Auch wenn es sich hier um den zweiten Block handelt, so kann es vorkommen, dass manche Items vor der *Construct Map* erstellt werden. Wichtig ist am Ende nur, dass die erstellten Items als Folgerungen aus der *Construct Map* zu sehen sind (vgl. Wilson & Mark, 2005).

2.2.3 Outcome Space

Im *Outcome Space* werden die zuvor festgelegten Items nun kategorisiert und mit einem Wert versehen. Zum Beispiel können Items als „Wahr oder Falsch“, „Multiple Choice“ oder „Open End“ deklariert werden. Der Wert, der einem Item zugeteilt wird, hängt auch von der zuvor festgelegten Kategorie ab. Bei einem einfachen *Wahr oder Falsch* Item, können die zwei Optionen jeweils mit 0 und 1 gewertet werden, bei Items, die mehrere Antwortmöglichkeiten zulassen, muss nun absteigend von sehr zustimmend bis hin zu sehr ablehnend auch der Wert der Antwort abnehmen. Dies wird dann im nächsten Block verwendet (vgl. Wilson & Mark, 2005).

2.2.4 Measurement Model

Im *Measurement Model* muss nun entschieden werden, welcher Messansatz für das Messtool verwendet wird. Hier gibt es im Groben zwei Möglichkeiten. Bei der einen werden Items, die alle dieselbe Ausprägungsrichtung beinhalten, mit entsprechenden Punkten versehen. Am Ende der Befragung werden diese Punkte dann addiert. Je höher die Punkteanzahl, desto höher die entsprechende Merkmalsausprägung. Die andere Möglichkeit bietet einer Testperson unterschiedliche Aussagen, denen zugestimmt oder die abgelehnt werden können. Die Aussagen bilden beide Ausprägungsrichtungen der Charakteristika des Aspektes ab. Einer Zustimmung wird zum Beispiel „1“ und einer Ablehnung „0“ zugeordnet. So ergibt sich wieder durch einen hohen Score am Ende eine hohe Merkmalsausprägung (vgl. Korner, 2014; Wilson & Mark, 2005).

3 Methodik - Zeitverzögerter Test

Bei einem zeitverzögerten Test (oder auch *Follow Up*-Test) handelt es sich um eine Messmethode für „Vergleichsstudien zur Wirkung von Unterrichtsmedien“ (Krüger et al., 2014, S.67). Hierbei will man langfristige Effekte nachweisen, indem man nach einem gewissen zeitlichen Abstand zur Intervention einen Test durchführt, der im besten Fall bereits in dieser (oder ähnlicher) Form als Test (unmittelbar nach der Intervention) von den Schüler*innen bearbeitet wurde (vgl. Krüger et al., 2014, S.71).

Der Test sollte allen Schüler*innen vorgelegt werden, die entweder an der Unterrichtskonzeption von Hannah Loidl oder von Rupert Bärenthaler-Pachner teilgenommen haben, um die Ergebnisse mit denen der beiden Kolleg*innen vergleichen zu können. Die Auswertung der Testergebnisse sind im Abschnitt 4 (Datenauswertung) zu finden

3.1 Erstellen des Tests

Loidl und Bärenthaler-Pachner führten am Ende ihrer Konzeptionen jeweils einen Test durch. Beide Tests beinhalteten geschlossene und offene Antwortformate, wobei der von Loidl überwiegend aus offenen und der von Bärenthaler-Pachner zum Großteil aus geschlossenen Elementen bestand. Zu beachten war beim Erstellen des Tests, dass die Konzeption von Loidl für Unterstufe konzipiert war und daher die rechnerische Komplexität dementsprechend niedriger war, als die bei Bärenthaler-Pachner.

Aus dem Test von Loidl wurden die ersten beiden Beispiele (Aufgabe 6. im zeitverzögerten Test) entnommen. Diese wurde, bis auf eine kleine Veränderung, genauso aus dem Test übernommen. Die Änderung rührte daher, dass die Aufgabenstellungen von Loidl so aufgebaut waren, dass immer eine Seite nach der nächsten bearbeitet wurde. Die Angabe in Beispiel 2 gab also einen Hinweis auf die Antwort in Beispiel 1, was bei dem nun erstellten verzögerten Test ein Problem darstellte, da die Art der Testbearbeitung aus praktischen Gründen nicht übernommen wurde.

Daher wurde anstatt „Martin führt nun das Experiment mit der Stoppuhr seines Handys durch“ (Loidl, 2021, S.121) folgende Formulierung verwendet: „Martin führt nun das Experiment mit verbesserten Methoden durch“.

Aus der schriftliche Informationsfeststellung von Bärenthaler-Pachner wurden die ersten vier Beispiele entnommen und Beispiel 1 (Aufgabe 1.) unverändert, die anderen drei stark verändert in den *Follow Up*-Test übernommen. Dem Beispiel 2 (Aufgabe 3.) von Bärenthaler-Pachner wurde die Aussage „Die Firma, die Längenmessgeräte herstellt, ändert das Design der Anzeige eines Messgerätes.“, welche direkt einem Arbeitsblatt von Bärenthaler-Pachner entnommen wurde, hinzugefügt. Beispiel 3 (Aufgabe 2.) wurde mithil-

fe eines anderen Antwortformats eingefügt da das bestehende Format als unübersichtlich befunden wurde. Am stärksten verändert wurde Beispiel 4 (Aufgabe 4.). Hier wurde, wie bei Aufgabe 3. auf ein *n aus 5*-Antwortformat zurückgegriffen und zwei Optionen wurden hinzugefügt.

Zusätzlich wurde eine Aufgabe zur Gänze neu erstellt. Gemeinsam mit den zuvor beschriebenen Aufgaben konnten alle Lernziele von Loidl und Bärenthaler-Pachner mit dem zeitverzögerten Test abgedeckt werden. Im Folgenden ist der vollständige Test inklusive Musterlösung und abgedeckter Lernziele zu finden.

Auf den Test können insgesamt 15 Punkte erreicht werden. Die Aufgaben 1 bis 6a wurden alle mit jeweils zwei Punkten und Aufgabe 6b mit drei Punkten gewichtet. Bei den Aufgaben 1 und 5 gibt es jeweils nur eine richtige Antwort, weshalb hier entweder null oder zwei Punkte vergeben werden. Da die Antworten bei Aufgabe 3 richtig und teilrichtig sein können, werde hier entweder null, ein oder zwei Punkte vergeben. Bei den Aufgaben 3 und 4 sind jeweils zwei der fünf möglichen Antworten richtig. Hier werden zwei Punkte vergeben, wenn beide richtigen gekreuzt werden und ein Punkt, wenn nur eine richtige gekreuzt wird. Sobald eine der falschen Antworten ausgewählt werden, werden null Punkte vergeben. Bei Aufgabe 6a wird ein Punkt für die richtige Einschätzung der Vertrauenswürdigkeit gegeben, wobei hier als richtig die ersten drei Skalenteile gewertet werden. Der zweite Punkt wird für richtige Verbesserungsvorschläge gegeben, wobei hier, sobald ein falscher Vorschlag genannt wird, sofort null Punkte vergeben werden. Zuletzt teilen sich die drei Punkte bei Aufgabe 6b auf die drei Teilaufgaben auf. Also ein Punkt für die richtige Einschätzung, ein Punkt für die richtige Begründung und ein Punkt für richtige Verbesserungsvorschläge.

Welche Lernziele mit welchen Items abgedeckt sind und wie viele Punkte jeweils dafür gegeben werden, ist überblicksmäßig noch in Tabelle1 zu finden.

Item-Nummer	Lernziele	erreichbare Punkte
Item 1	SEKI: Die Vertrauenswürdigkeit verschiedener Messergebnisse einschätzen und vergleichen können. SEKII: Lernende können Messungen auf Basis ihrer Vertrauenswürdigkeit bewerten. Dabei können sie verschiedene Messungen nach ihrer Messunsicherheit ordnen und Vergleiche bezüglich der Vertrauenswürdigkeit ziehen.	0/2 Punkte
Item 2	SEKI: Ein Ergebnis vollständig (Mittelwert + Streuung) angeben können. SEKII: Lernende geben Messergebnisse richtig an (Mittelwert $\pm$ Messunsicherheit) und unterscheiden zwischen korrekten und falschen Angaben wissenschaftlicher Messwerte.	0/1/2 Punkte
Item 3	SEKI: Verschiedene Quellen für Unsicherheiten aufzählen und sie bei der Beurteilung der Vertrauenswürdigkeit miteinbeziehen können. SEKII: Lernende unterscheiden Messunsicherheiten nach Typ- A und Typ-B	0/1/2 Punkte
Item 4	SEKI: Verschiedene Quellen für Unsicherheiten aufzählen und sie bei der Beurteilung der Vertrauenswürdigkeit miteinbeziehen können. SEKII: Lernende beschreiben Schritte bei der Herstellung und Verwendung eines Messgeräts, die zu Typ-B-Messunsicherheiten führen.	0/1/2 Punkte
Item 5	SEKI: Die Vertrauenswürdigkeit verschiedener Messergebnisse einschätzen und vergleichen können. SEKII: Lernende können Messungen auf Basis ihrer Vertrauenswürdigkeit bewerten. Dabei können sie verschiedene Messungen nach ihrer Messunsicherheit ordnen und Vergleiche bezüglich der Vertrauenswürdigkeit ziehen.	0/2 Punkte
Item 6a und	SEKI: Die Vertrauenswürdigkeit verschiedener Messergebnisse einschätzen und vergleichen können. SEKII: Lernende können Messungen auf Basis ihrer Vertrauenswürdigkeit bewerten. Dabei können sie verschiedene Messungen nach ihrer Messunsicherheit ordnen und Vergleiche bezüglich der Vertrauenswürdigkeit ziehen.	0/1/2 Punkte
Item 6b	SEKII: Verschiedene Quellen für Unsicherheiten aufzählen und sie bei der Beurteilung der Vertrauenswürdigkeit miteinbeziehen können.	0/1/2/3 Punkte

Tabelle 1: Übersicht der Lernziele der Items nach Sekundarstufe und Punkteverteilung.

Für die Verbesserungsvorschläge werden Kategorien verwendet, die von Loidl (2021) teilweise übernommen und gemeinsam mit Brasseur (2024) überarbeitet wurden. Im Folgenden sind diese aufgelistet und näher erklärt, wobei die Kategorie K-1.3 nicht im Zuge der offenen Aufgaben festgestellt wird, sondern mittels den beiden anderen Teilaufgaben in Aufgabe 6.

K-1.1 Messreihe anfertigen; Mittelwert, Spannweite und/oder andere statistische Größen berechnen

K-1.2 Streuungsmaße untersuchen

K-1.3 Vertrauenswürdigkeit beurteilen

K-1.4 Ausprobieren und Optimieren

K-2.1 Messunsicherheiten diskutieren

K-2.2 Streuung von Messwerten (Typ A-MUS)

K-2.3 Genauigkeit des Messgeräts (Typ B-MUS)

K-3.1 Handlung ohne Ziel

K-3.2 Fehlerhafte Handlung

Aufgabe 0.

- ☐ Es **wurde** im Unterricht an dem Konzept *Messunsicherheiten* weitergearbeitet.
- ☐ Es **wurde nicht** im Unterricht an dem Konzept *Messunsicherheiten* weitergearbeitet.

Aufgabe 1.

Ordne die folgenden Zeitmessgeräte nach ihrer Vertrauenswürdigkeit, beginnend mit der geringsten Vertrauenswürdigkeit.

- A Stoppuhr auf einem Smartphone ($\pm 0,01$ s)
- B Armbanduhr mit einem Sekundenzeiger
- C Digitale Stoppuhr, die auf Zehntelsekunden genau misst ($\pm 0,1$ s)
- D Atomuhr die auf 10^{-18} s (Trillionstel Sekunde) genau misst.

Kreuze die richtige Reihenfolge an. [1 aus 4]

A-C-D-B	☐
B-C-A-D	☒
A-B-C-D	☐
C-B-A-D	☐

→ SuS können Messungen auf Basis ihrer Vertrauenswürdigkeit bewerten. Dabei können sie verschiedene Messungen nach ihrer Messunsicherheit ordnen und Vergleiche bezüglich der Vertrauenswürdigkeit ziehen.

Aufgabe 2.

Vier Schüler*innen haben folgende Aufgabe bearbeitet:

Markiere in der untenstehenden Tabelle die beiden Ergebnisse, die wissenschaftlich korrekt angegeben sind.

Darstellung des Ergebnisses	
$v = 130$ km/h	A
$A = (34,0234 \pm 5,2353444223)$ m ²	B
$M = (40,1 \pm 0,1)$ kg	C
$r = 52$ cm $\pm$ 1 cm	D

Kreuze an, welche der 4 unterschiedlichen Lösungen der Schüler*innen korrekt ist. [1 aus 4]

A, C	☐
B, C	☐
C, D	☒
B, D	☐

→ SuS geben Messergebnisse richtig an (Mittelwert $\pm$ Messunsicherheit) und unterscheiden zwischen korrekten und falschen Angaben wissenschaftlicher Messwerte.

Aufgabe 3.

Welche der folgenden angeführten Beispiele tragen zur Messunsicherheit bei?
 Kreuze die zutreffende(n) Aussage(n) an. *[n aus 5]*

Beim Skalieren wird zufälligerweise eine Markierung zu weit rechts eingezeichnet	☒
Der Beifahrer in einem Auto soll eine Geschwindigkeit ablesen, die sich genau zwischen zwei Skalenteilen des Tachometers befindet.	☒
Die Firma, die Eichnormale für eine Masse von 10 kg herstellt, ändert deren Form von Kugeln auf Würfel.	☐
Durch Zufall wird eine Eichlänge etwas zu kurz übertragen.	☒
Die Firma, die Längenmessgeräte herstellt, ändert das Design der Anzeige eines Messgerätes.	☐

→ SuS beschreiben Schritte bei der Herstellung und Verwendung eines Messgeräts, die zu Typ-B-Messunsicherheiten führen.

Aufgabe 4.

Welche der folgenden Beispiele gehören zu der Messunsicherheit, die durch die Unsicherheit des Messgerätes beeinflusst wird? (Typ-B-Messunsicherheit)
 Kreuze die zutreffende(n) Aussage(n) an. *[n aus 5]*

Auf einer Waage steht $\pm 0,5$ kg.	☒
Berechnung der Standardabweichung des Mittelwertes.	☐
Berechnung der Spannweite.	☐
Eine Stoppuhr kann auf 1,5 Sekunden genau messen.	☒
Die Abstände der Markierungen auf einem Lineal sind nicht alle gleich weit voneinander entfernt.	☐

→ SuS können die Messunsicherheiten Typ-B und Typ-A voneinander unterscheiden und entsprechenden Unsicherheiten zuordnen.

Aufgabe 5.

Welches der folgenden Ergebnisse ist das Vertrauenswürdigste?
 Kreuze die zutreffende Aussage an. *[1 aus 5]*

$m = (0,340 \pm 0,001)$ kg	☐
$m = (340,0 \pm 0,1)$ g	☒
$m = (340 \pm 10)$ g	☐
$m = (0,34 \pm 0,01)$ kg	☐
$m = (340 \pm 1)$ g	☐

→ SuS können Messungen auf Basis ihrer Vertrauenswürdigkeit bewerten. Dabei können sie verschiedene Messungen nach ihrer Messunsicherheit ordnen und Vergleiche bezüglich der Vertrauenswürdigkeit ziehen.

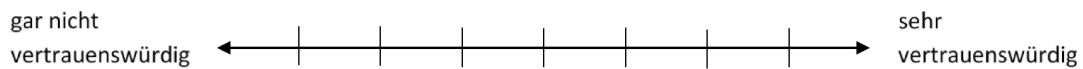
Aufgabe 6.

Martins Klasse soll im Physikunterricht folgendes Experiment durchführen:

Lasse einen Tennisball von zwei Metern Höhe zu Boden fallen und stoppe die Zeit.
Wie lange braucht der Ball, bis er am Boden aufprallt?

- a) Martin lässt den Ball aus zwei Metern Höhe fallen und stoppt die Zeit mit dem Sekundenzeiger seiner Armbanduhr. Die gestoppte Zeit beträgt **2 s**.

Wie vertrauenswürdig ist die Messung von Martin deiner Meinung nach?



Einer der ersten drei Skalenteile

Was könnte Martin an seiner Messung verbessern?

- 1) Genauigkeit des Messgeräts erhöhen (Lichtschraken, genauere (Stopp-)Uhr,

Videoaufzeichnungen, Sensoren,...) → Genauigkeit des Messgeräts (K-2.3)

- 2) Öfter messen und einen Durchschnitt ausrechnen. → Messreihe und Mittelwert (K-1.1/K-2.2)

Genauer stoppen, schneller reagieren, ... → Streuung der Messwerte (Typ A) (K-1.2/K-2.2)

- 3) Mit zweiter Person arbeiten, richtige Wahl der Perspektive, Verwenden eines Sessels,...

→ Verbesserung der Versuchsanordnung (K-1.4)

- b) Martin führt nun das Experiment mit verbesserten Methoden durch und vergleicht seine neu gemessene Zeit von **0,67 s** mit der gemessenen Zeit seiner Sitznachbarin Lisa. Lisa hat **0,90 s** gestoppt. Beide haben das Experiment unter denselben Bedingungen durchgeführt.

Wer von den beiden hat das vertrauenswürdiger Ergebnis?

- ☐ Lisa hat das vertrauenswürdiger Ergebnis.
☐ Martin hat das vertrauenswürdiger Ergebnis.
☒ Das kann ich so nicht beantworten.

Begründe bitte deine Antwort:

Die einmalige Messung der Zeit liefert kein aussagekräftiges Ergebnis.

Messunsicherheit ist nicht angegeben.

Könnten die beiden noch etwas an ihrem Experiment verbessern? Wenn ja, was zum Beispiel?

Siehe Item a)

→ SuS können verschiedene Quellen für Unsicherheiten aufzählen und sie bei der Beurteilung der Vertrauenswürdigkeit miteinbeziehen.

→ SuS können die Vertrauenswürdigkeit verschiedener Messergebnisse einschätzen und vergleichen.

3.2 Durchführung des Tests

Eine der größten Schwierigkeiten des gesamten Forschungsprojekts stellte das Auffinden der Schüler*innen, die damals an den Unterrichtskonzeptionen teilnahmen, dar. Die Durchführung bei Loidl lag bereits 2 Jahre zurück und fand damals in einer 4.Klasse der SEKI einer AHS statt. Das bedeutete, dass sich die Schüler*innen teilweise nicht mehr in der Schule befanden und auf unterschiedliche Klassen aufgeteilt waren. Insgesamt konnten von Loidl zwölf Schüler*innen relokalisiert werden. Diese befanden sich nun alle in der 2.Oberstufe (10.Schulstufe).

Bei Bärenthaler-Pachner lag die Durchführung ca. ein Jahr zurück. Hier gestaltete sich zwar das Auffinden der Schüler*innen leichter, jedoch befand sich eine Klasse in Osttirol und mit den aktuellen Physiklehrer*innen der anderen Klassen gestaltete sich die Kommunikation als herausfordernd. Bis zum Schluss konnten 29 Schüler*innen von Bärenthaler-Pachner erreicht werden. Diese befanden sich nun in der 4.Oberstufe (12.Schulstufe)

4 Methodik - Fragebogen zur Vertrauenswürdigkeit

Das Herzstück der Forschungsarbeit bildet der Fragebogen zur *Vertrauenswürdigkeit*. Da der Begriff in den beiden Unterrichtskonzeptionen von Loidl und Bärenthaler-Pachner eingebaut wurde, stellte sich die Frage, wie gut dieser von den Schüler*innen, insbesondere der SEKI, angenommen und verstanden wurde. Diese Frage sollte mit dem Fragebogen beantwortet werden.

Die Besonderheit an der Erarbeitung dieses Fragebogens bestand darin, dass eine Forschungsgruppe aus insgesamt fünf Personen gebildet wurde, bestehend aus vier Bachelorstudierende, die ihre Bachelorarbeit bei Clemens Nagel verfassten und der Verfasserin. Die beiden Ziele dieser Zusammenarbeit waren, dass die vier Studierenden einen Einblick in eine größere Forschungsarbeit bekommen konnten und dass die Forschungsarbeit, durch die große Unterstützung, in einem Semester abgeschlossen werden konnte.

4.1 Entwicklung des Fragebogens

Die Erstellung des Fragebogens bestand aus mehreren Phasen, die die vier *Building Blocks* (Wilson, 2004) in sich abdeckten. Zuerst wurden acht unterschiedliche Dimensionen von Vertrauenswürdigkeit festgelegt und auf Basis dieser und den dazugehörigen Construct Maps wurden erste Items erstellt. Auf Basis eines Experteninterviews (Helfferich, 2014) mit Clemens Nagel wurde dann eine Überarbeitung angefertigt. Dieser Entwurf wurde dann am AECC Physik präsentiert und das Feedback zeigte, dass dies kein adäquater Fragebogen für Zielgruppe und Forschungsfrage darstellte.

Um also einen geeigneten Fragebogen zu entwickeln, wurde zunächst der Alltagsbezug

von Schüler:innen sowie ihr grobes Verständnis des Begriffs Vertrauenswürdigkeit untersucht. Dafür wurden Brainstormings (Altrichter et al., 2018) in unterschiedlichen Klassen innerhalb von zwei Wochen durchgeführt. Mit den erhaltenen Informationen wurden entsprechende Items entwickelt und zu einem vorläufigen Fragebogen zusammengefasst. Dieser neue Entwurf wurde einem Pretesting unterzogen, welches aus sogenannten „Think Aloud“ (Lautes Denken) Interviews (Krüger et al., 2014) mit Schüler*innen, die den Fragebogen bearbeiteten, bestand. Innerhalb von vier Wochen fand die Durchführung, Auswertung und Prüfung der Reliabilität der Ergebnisse dieser Interviews mittels der Inter-Coder-Reliabilität (O'Connor & Joffe, 2020) statt. Im Anschluss daran wurde der Fragebogen in seine (vorläufig) finale Form gebracht.

4.1.1 Aufsuchen der Dimensionen und erster Fragebogenentwurf

Der erste Schritt in der Erstellung des Fragebogens war es herauszufinden, in welche Dimensionen (*Constructs*) sich der Begriff *Vertrauenswürdigkeit* einteilen lässt. Hierfür wurde Luhmanns *Vertrauen - Ein Mechanismus der Reduktion sozialer Komplexität* zugezogen. Auf Basis dieser Literatur wurden folgende Konstrukte und Dimensionen erarbeitet:

- 1) **Informationsweitergabe:** Vertrauenswürdigkeit hat immer etwas mit Informationsweitergabe zu tun. — Vertrauenswürdigkeit hat nie etwas mit Informationsweitergabe zu tun.
- 2) **Wissenschaft-Alltag:** Der Begriff der Vertrauenswürdigkeit in der Wissenschaft bedeutet dasselbe wie im Alltag. — Der Begriff der Vertrauenswürdigkeit in der Wissenschaft unterscheidet sich grundlegend von dem im Alltag.
- 3) **Positiv-negativ:** Der Begriff der Vertrauenswürdigkeit beinhaltet ausschließlich positive Aspekte. — Der Begriff der Vertrauenswürdigkeit beinhaltet ausschließlich negative Aspekte.
- 4) **Menschlich-künstlich:** Vertrauenswürdigkeit ist etwas ausschließlich Menschliches. — Vertrauenswürdigkeit ist etwas ausschließlich Künstliches/Maschinelles.
- 5) **Subjektiv - Objektiv:** Vertrauenswürdigkeit ist etwas ausschließlich Subjektives. — Vertrauenswürdigkeit ist etwas ausschließlich Objektives.
- 6) **Handelnde - Handlungen:** Vertrauenswürdigkeit wird ausschließlich von den Handelnden bestimmt. — Vertrauenswürdigkeit wird ausschließlich von den Handlungen bestimmt.
- 7) **Wiederherstellung:** Vertrauenswürdigkeit kann immer, nachdem es zerstört wurde, wiederhergestellt werden. — Vertrauenswürdigkeit kann nie, nachdem es zerstört wurde, wiederhergestellt werden.

- 8) **(Un)verzichtbar:** Vertrauenswürdigkeit ist ein unverzichtbares Attribut. — Vertrauenswürdigkeit ist ein verzichtbares Attribut.

Mit diesen Dimensionen wurde versucht, den Begriff *Vertrauenswürdigkeit* in allen Facetten abzubilden. Auf Basis der Dimensionen wurden erste Items für den Fragebogen erstellt. Diese sollten die unterschiedlichen Abstufungen der Dimensionen abdecken und damit entsprechend dem *Item Design* Block entsprechen. Diese Items waren äußerst allgemein und ohne Alltagsbezug formuliert. Im Folgenden sind ein paar Beispiele zu finden:

- Wenn eine Quelle vertrauenswürdig ist, dann glaube ich jede Information, die von ihr kommt.
- Ob eine Quelle vertrauenswürdig ist, hängt hauptsächlich von der Quelle selbst ab.
- Wenn eine andere, mir nahestehende, Person etwas für vertrauenswürdig hält, dann gilt dies automatisch für mich.
- Wenn eine Person für mich vertrauenswürdig ist, dann liegt das alleine an ihren Handlungen.

Diese und mehr als 60 weitere solcher Items wurde im Zuge des ersten Fragebogenentwurfs erstellt. Die vollständige Auflistung der Items ist im Anhang unter A zu finden.

4.1.2 Experteninterview

Der nächste Teil bestand aus einem Interview mit Clemens Nagel. Ein Experteninterview ist eine qualitative Forschungsmethode, bei der gezielt Fachleute oder Personen mit spezifischem Wissen und Erfahrung in einem bestimmten Bereich befragt werden. Ziel ist es, tiefgehende und fundierte Einblicke zu gewinnen, die über allgemeines Wissen hinausgehen. Diese Interviews bieten detaillierte Informationen, die durch die Expertise und das praktische Wissen der Interviewten bereichert werden. Sie sind besonders nützlich in Bereichen, in denen spezifisches Fachwissen erforderlich ist, und helfen, komplexe Themen aus der Perspektive von Fachleuten zu verstehen und zu analysieren (Helfferich, 2014).

Nagel hatte zuvor den Fragebogenentwurf eingesehen und entsprechend den Kommentaren wurden dann die bestehenden Items gemeinsam untersucht und diskutiert. Hierbei stellte sich heraus, dass die Items zur Gänze neu formuliert werden mussten, da der fehlende Alltagsbezug den Fragebogen nicht bearbeitbar machte für die Schüler*innen.

Auf Basis des Interviews wurden neue Items erstellt, die auch das Thema *Messungen* behandelten, und eine Rubrik von Synonym- und Gegenteilabfragen wurde ergänzt. Hierbei sollten die Schüler*innen unterschiedliche Begriffe kategorisieren, ob diese gleichbedeutend,

verschieden oder nichts mit *Vertrauenswürdigkeit* zu tun hatten.³ Diese Version des Fragebogens ist im Anhang unter B zu finden.

4.1.3 Experten-Diskussion

Dieser Entwurf wurde im Rahmen der Lehrveranstaltung *Fachdidaktische Entwicklungsforschung - aktuelle Entwicklungen* am AECC Physik (Austrian Educational Competence Centre Physics) präsentiert und anschließend diskutiert.

Hierbei handelt es sich um eine Expertendiskussion, welche ein strukturiertes Gesprächsformat ist, bei dem mehrere Fachleute aus einem bestimmten Bereich zusammenkommen, um über spezifische Themen oder Fragestellungen zu debattieren und ihre unterschiedlichen Perspektiven und Erkenntnisse auszutauschen. Dieses Format ermöglicht es, komplexe Probleme umfassend zu beleuchten, verschiedene Standpunkte zu berücksichtigen und innovative Lösungsansätze zu entwickeln. Das Feedback zeigte, dass das Ausmaß an Dimensionen den Rahmen einer Masterarbeit sprengte und somit die Forschungsarbeit grundlegend überarbeitet werden musste (Bogner, 2002).

4.1.4 Pretesting 1 - Brainstorming in sechs Klassen

In Zusammenarbeit mit der Forschungsgruppe wurde das Pretesting 1, das Brainstorming, in insgesamt sechs Klassen an drei verschiedenen Schulen von vier Mitgliedern der Forschungsgruppe durchgeführt. Dafür wurden den Schüler*innen leere Karten ausgeteilt, auf die sie in Einzelarbeit alles aufschreiben sollten, was ihnen zu der Frage „Was ist Vertrauenswürdigkeit für dich?“ einfällt. Diese Kärtchen wurden nach ca. drei Minuten Arbeitszeit wieder eingesammelt und anschließend auf der Tafel in Gruppen, ausgehend von Schüler*innen-Assoziationen, sortiert. Letzteres wurde durchgeführt um anschließend mögliche Dimensionen des Begriffs aus Schüler*innensicht besser definieren zu können.

Mit einem Brainstorming können viele unterschiedliche Ansätze von vielen Schüler*innen gleichzeitig eingeholt werden, dabei wird es als „kreative Methode“ bezeichnet (Altrichter et al., 2018).

Pretestings, auch als Vorversuche oder Pilotprojekte bezeichnet, sind methodische Vorabtests, die durchgeführt werden, um Unterrichtsmaterialien, -methoden oder -strategien in kleinerem Rahmen zu erproben, bevor sie im Großen eingesetzt werden. Das Ziel von Pretestings ist es, mögliche Schwachstellen, Missverständnisse oder praktische Probleme frühzeitig zu identifizieren und zu beheben. Dadurch wird die Effektivität und Effizienz des Forschungsinstruments verbessert. Pretestings ermöglichen es, Rückmeldungen zu sammeln und Ansätze auf Basis empirischer Daten zu optimieren (Altrichter et al., 2018).

³Zu diesem Zeitpunkt wurde auch die Forschungsgruppe mit den vier Kolleg*innen gebildet.

Aufgrund der Resultate wurde in einigen Klassen anschließend noch ein zweiter Durchgang mit der Frage „*Wen oder was hältst du für vertrauenswürdig*“ abgehalten. Diese Frage wurde gewählt, da im ersten Durchgang fast ausschließlich Personen genannt wurden. So sollte überprüft werden, ob die Schüler*innen auch Objekt mit dem Begriff in Verbindung bringen und welche diese sind. Zusätzlich wurde in manchen Klassen gefragt, was die Schüler*innen unter *Glaubwürdigkeit* verstehen, da zuvor mehrere Male die Begriffe *vertrauen* und *glauben* in gleichen Kontexten verwendet wurden. Es wurde also überprüft, ob die Schüler*innen den Begriff *Glaubwürdigkeit* eher mit der Informationsweitergabe von Fachwissen und nicht nur mit Personen und den damit zusammenhängenden Geheimnissen in Verbindung brachten. Bei beiden Zusatzfragen wurden die Kärtchen ebenfalls an der Tafel in Gruppen sortiert und mit den Schüler*innen diskutiert.

4.1.5 Überarbeitung in Gruppendiskussion - Reduktion der Dimensionen

Auf Basis des *Pretesting 1* wurden die acht Dimensionen auf die Dimension *Menschlich - künstlich* reduziert und diese dafür aufgespalten in die Subdimensionen *Menschlich (Person)* - *Menschlich (digital)* und *Künstlich (dinghaft)* - *Künstlich (nicht dinghaft)*. Diese vier Subdimensionen ergaben sich auf Basis der Gruppen, in die die Begriffe der Brainstormings eingeteilt wurden. Diese sind in der nachfolgenden Tabelle 2 gelistet.

Menschlich (Person)	Menschlich (digital)	Künstlich (dinghaft)	Künstlich (nicht dinghaft)
Eltern Lehrpersonen Freunde Selbstvertrauen	Soziale Medien: Instagram YouTube Lernvideos TikTok	Lineal Thermometer Maschinen Schulbücher Wörterbücher	Internet Wikipedia Chat GPT

Tabelle 2: Die vier erhaltenen Subdimensionen zu dem Konstrukt „Menschlich - künstlich“

Auf Basis dieser Subdimensionen wurde in Form einer Gruppendiskussion der zweite Fragebogenentwurf erstellt, welcher im Anhang unter C zu finden ist.

Gruppendiskussionen, auch bekannt als Fokusgruppen oder „focus groups“, sind strukturierte Gespräche, die darauf abzielen, Einstellungen zu einem bestimmten Thema zu erfassen, das durch das Forschungsinteresse definiert ist. Sie finden in einer offenen und freundlichen Atmosphäre statt und fördern Kommunikationsprozesse, die einem alltäglichen Gespräch ähneln. Dabei steht nicht nur der Austausch von Argumenten im Vordergrund, sondern es wird auch erzählt, erinnert und sich gegenseitig ergänzt. Die Interaktion der Gruppenmitglieder ist entscheidend, was die Methode über die einfache Befragung mehrerer Personen hinaushebt.

Gruppendiskussionen sind in der Regel qualitativ ausgerichtet und streben nach Natürlichkeit, Kommunikativität und Offenheit in der Erhebungssituation. Durch die Interaktion

der Teilnehmer untereinander und die damit verbundene Beeinflussung sind sie realistischer und näher an natürlichen Gesprächssituationen als beispielsweise Einzelinterviews. Aufgrund des diskursiven Austauschs können sich während der Diskussion auch Argumente verändern.

Diese Methode dient entweder dazu, Einzelmeinungen in validierter Form oder Gruppenmeinungen losgelöst vom Individuum zu erheben. Sie können entweder bereits vorhandene Meinungen oder den Prozess der Meinungsbildung und -verhandlung untersuchen. Individuelle Einstellungen oder Erzählungen subjektiver Erfahrungen werden dagegen in Gruppendiskussionen nicht erfasst (Vogl, 2022).

Die ersten elf Items sind sogenannte Gegenteil- bzw. Synonymabfrage Items, wobei es sich bei Item 4 und 8 (Faulheit und Langeweile) um Kontrollfragen handelt. Die Antwortmöglichkeiten sind jeweils „sehr ähnlich“, „sehr verschieden“ und „haben nichts miteinander zu tun“. Diese sollen dabei helfen, festzustellen, ob die Schüler*innen die Items gewissenhaft bearbeiten.

Items 1, 6, 7 und 9 sind als Gegenteile angelegt, während es sich bei 2, 3, 5 und 10 um Synonyme handelt. Des weiteren wurde Item 11 (geringe Messunsicherheit) eingebaut, um sich bereits ein Bild davon zu machen, welchen Zusammenhang die Schüler*innen der SEKI zwischen *Vertrauenswürdigkeit* und *geringer Messunsicherheit* sehen. Diese Einteilung ist in der anschließenden Tabelle 3 überblicksmäßig angeführt.

Kontrollfragen	4 Faulheit und Vertrauenswürdigkeit sind für mich. . .
	8 Langeweile und Vertrauenswürdigkeit sind für mich. . .
Gegenteilsabfrage	1 Unsicherheit und Vertrauenswürdigkeit sind für mich. . .
	6 Unehrlichkeit und Vertrauenswürdigkeit sind für mich. . .
	7 Unglaubwürdigkeit und Vertrauenswürdigkeit sind für mich. . .
	9 Zwielfichtigkeit und Vertrauenswürdigkeit sind für mich. . .
Synonymabfrage	2 Richtigkeit und Vertrauenswürdigkeit sind für mich. . .
	3 Glaubwürdigkeit und Vertrauenswürdigkeit sind für mich. . .
	5 Verlässlichkeit und Vertrauenswürdigkeit sind für mich. . .
	10 Aufrichtigkeit und Vertrauenswürdigkeit sind für mich. . .
Messbezug	11 Geringe Messunsicherheit und Vertrauenswürdigkeit sind für mich. . .

Tabelle 3: Einteilung der entwickelten Gegenteil- und Synonymabfrage Items

Anschließend wurden zu den in Tabelle 2 gelisteten Begriffen und Subdimensionen die Items 12-22 und 31-34 formuliert. Dafür wurde ein einheitlicher Wortlaut gewählt, um im Vorhinein Verständnisverfälschungen durch unterschiedliche Ausdrucksweisen ausschließen zu können. Das Item 31 wurde außerdem so formuliert, dass ein Begriff aus der Subdimension „Menschlich (dinghaft)“ verwendet wurde und auch eindeutig hervorgeht, dass der

Mensch hinter dem sozialen Medium in seiner Vertrauenswürdigkeit eingestuft werden soll und nicht das Medium selbst. Die Items 18, 32 und 34 wurden bewusst als negativ polierte Items der Subdimension „Menschlich (Person)“ formuliert. Die Antworten der Items 12-34 sind in Form einer fünfstufigen Likert-Skala möglich von „ich stimme sehr zu“ bis „ich stimme gar nicht zu“. Diese Items sind in der Tabelle 4 zu finden.

Menschlich (Person)	13 Ich halte Informationen von meiner Lehrperson für vertrauenswürdig.
	14 Ich halte Informationen von meinen Eltern für vertrauenswürdig.
	15 Ich halte Informationen von meinem/meiner Nachhilfelehrer*in für vertrauenswürdig.
negativ poliert zu Menschlich (Person)	18 Ich halte Informationen von unbekannten Anrufern für vertrauenswürdig.
	32 Mein Großvater kann mir nicht die 9 Bundesländer Österreichs nennen. Ich vertraue ihm trotzdem.
	34 Ich halte fremde Personen, die mir am Bahnsteig irgendwelche Waren anbieten wollen, für vertrauenswürdig.
Menschlich (dinghaft)	17 Ich halte Informationen von YouTube-Lernvideos für vertrauenswürdig.
	20 Ich halte Informationen auf Sozialen Medien (Instagram, TikTok,...) für vertrauenswürdig.
	31 Ein YouTube-Kanal dem ich vertraue, wechselt den Content-Creator (Autor). Ich vertraue dem Kanal weiterhin.
Künstlich (dinghaft)	12 Ich halte Informationen in meinem Schulbuch für vertrauenswürdig.
	21 Ich halte die Ergebnisse des Taschenrechners für vertrauenswürdig.
	22 Ich halte die Ergebnisse im Lösungsheft für vertrauenswürdig.
Künstlich (nicht dinghaft)	16 Ich halte Informationen von Wikipedia für vertrauenswürdig.
	19 Ich halte Informationen aus dem Internet für vertrauenswürdig.
	33 Die Antworten die mir ChatGPT (ChatBot) auf eine Frage gibt, halte ich (grundsätzlich) für vertrauenswürdig.

Tabelle 4: Entwickelte Fragen zu den erhaltenen Dimensionen

Mit den Items 23-30 sollte der Bezug zu Messungen im Alltag der Schüler*innen abgedeckt werden. Die Items 23-25 und 30 sind so formuliert, dass die Person hinter der Messung

die Einstufung der Vertrauenswürdigkeit beeinflussen soll. In den Items 23-25 wird dies nochmals verdeutlicht, indem drei gleiche Situationen mit unterschiedlichen Personen genutzt werden. Bei den Items 26-29 soll hingegen die Vertrauenswürdigkeit des Messgeräts eingeschätzt werden. Dieser Block ist in Tabelle 5 abgebildet.

messende Personen	23 Ein*e Freund*in und ich stoppen die Zeit bei einem Wettlauf. Ich vertraue meinem Messergebnis mehr als seinem/ihrer.
	24 Ein*e Wissenschaftler*in und ich stoppen die Zeit bei einem Wettlauf. Ich vertraue meinem Messergebnis mehr als seinem/ihrer.
	25 Eine Lehrperson und ich stoppen die Zeit bei einem Wettlauf. Ich vertraue meinem Messergebnis mehr als seinem/ihrer.
	30 Eine Schulärztin misst am Montag meine Körpergröße, das Messergebnis ist 159,0 cm. Am Dienstag misst sie nochmal. Nun ist das Ergebnis 158,5 cm. Ich halte dieses Messergebnis für vertrauenswürdig.
Messgerät	26 Ich wiege beim Backen das Mehl ab. Die Waage zeigt minus 10 Gramm an. Ich halte dieses Messergebnis für vertrauenswürdig.
	27 Ich wiege eine Kilopackung (1000 Gramm) Mehl auf einer Waage ab. Die Waage zeigt 999g an. Ich halte dieses Messergebnis für vertrauenswürdig.
	28 Mit einem Lineal kann ich vertrauenswürdig eine Länge von einem Zentimeter messen.
	29 Draußen schneit es. Das Außenthermometer zeigt 40°C an. Ich halte dieses Messergebnis für vertrauenswürdig.

Tabelle 5: Entwickelte Fragen mit deren Messbezug

4.1.6 Pretesting 2 - Think Aloud Interviews

Die Erhebung von gesprochenen Daten mittels der Lautes Denken-Methode ermöglicht einen Einblick in den Denkprozess und dynamische Aspekte der Problemlösung. Speziell deswegen bietet diese Methode für die „Testentwicklung [...] ein probates Mittel zur Prüfung der Aufgabengüte und Item-Validierung (Krüger et al., 2014, 182).“ Außerdem können diese Daten klare Verweise auf ungewollte Verständnisprobleme und Mehrdeutigkeiten herstellen. Diese Verweise tragen grundlegend zur Item Optimierung bei (Krüger et al., 2014).

Die Versuchspersonen müssen gemäß dem Forschungsinteresse ausgewählt werden und erhalten die Instruktion: bestmöglich alles, was ihnen durch den Kopf geht (ungeachtet ob es vollständige oder korrekte Gedanken sind) zu verbalisieren. Dafür sollen die Versuchspersonen während ihrer Antwort ununterbrochen sprechen. Unter anderem, weil dabei sehr schnell viele Informationen kommuniziert werden, ist es üblich, bei der Erhebung von verbalen Daten, eine Audiospur aufzunehmen und anschließend für die Weiterverarbeitung zu transkribieren. Die Daten werden in Folge häufig systematisch reduziert

und kategoriengeleitet analysiert. Für diese Analyse ist es notwendig eine Analyseeinheit zu definieren, also wie umfangreich ein Code ist. Diese Codes werden anschließend den vorab entwickelten Kategorien zugewiesen. Da die so analysierten Daten sehr prozessnahe gewonnen wurden, wird in der Regel davon ausgegangen, dass sie hoch valide sind. In anderen Worten: es wird davon ausgegangen, dass die Daten stark mit dem ihnen zugrunde liegenden Denkprozess in Verbindung stehen. Umstrittener hingegen ist die Reliabilität dieser Daten, also die zuverlässige Zuordnung dieser Daten in die Kategorien (Krüger et al., 2014).

In dieser Arbeit wird die Reliabilität mittels Intercoder Reliabilität Testung (siehe Abschnitt 4.2.1) überprüft. Im Rahmen dieser Arbeit wird mit 14 Schüler:innen ein Lautes Denken – Interview durchgeführt. Die Schüler:innen besuchen alle die Unterstufe einer allgemeinen höheren Schule in Österreich. Die Interviews werden von vier unterschiedlichen Personen geführt und aufgenommen und von fünf Personen (den vier Interviewer:innen und einer weiteren Person) transkribiert und codiert.

Die Interviews bestehen aus den 34 Items des Fragebogens 1 und Item-spezifischen Probing-Fragen. Die interviewten Schüler:innen sind aufgefordert, zu jedem Zeitpunkt laut mitzudenken. Die gesamte, auf eine Frage folgende Antwort, wird als Analyseeinheit definiert und entweder der Kategorie auffällig (bedeutet: möglicherweise Verbesserungspotential in der Item-Formulierung) oder unauffällig zugeordnet.

Mithilfe der als auffällig codierten Transkripten werden anschließend im Kollektiv die Items des Fragebogens 1 überarbeitet.

Auf Basis der Think Aloud Interviews wurden die Items 11, 23, 24, 25, 26, 27, 28, 30 und 31 geändert und die Items 9 und 32 gelöscht. Des weiteren wurden die Items 33 und 34 aufgrund der inhaltlichen Übereinstimmung weiter nach oben verschoben. Zu guter Letzt wurde ein neues Item bei der Gegenteil- und Synonymabfrage hinzugefügt.

Item 11 ursprünglich:	Geringe Messunsicherheit und Vertrauenswürdigkeit sind für mich:
Abgeändert:	Kleine Unsicherheit beim Messen und Vertrauenswürdigkeit...

Tabelle 6: Veränderung des Items 11

Bei Item 11 wurde einerseits der Begriff der Messunsicherheit umschrieben, da bei diesem (wie in den nachfolgenden Zitaten gezeigt) oftmals ein Unverständnis festgestellt werden konnte.

„Ich weiß jetzt nicht so genau, was geringe Messunsicherheit bedeutet ... hab ich noch nie gehört“ (11-N2)

„Was meinen Sie mit geringer Messunsicherheit?“ (11-B3)

Um Probleme mit der doppelten Verneinung sichtbar zu machen, wurde hier auch noch ein weiteres Item zum Vergleich hinzugefügt, mit der Formulierung „Große Unsicherheit beim Messen und Vertrauenswürdigkeit...“.

Andererseits wurde das Itemende abgeändert. Dies hatte seine Begründung darin, dass die ursprüngliche Endung nicht grammatikalisch korrekt mit allen ursprünglichen Antwortmöglichkeiten verbunden werden konnte. Deshalb wurden ebenfalls die Antwortmöglichkeiten geändert.

Ursprüngliche Antwortmöglichkeiten:

- sehr ähnlich
- sehr verschieden
- haben nichts miteinander zu tun

Neue Antwortmöglichkeiten:

- sind ähnlich
- sind gegenteilig
- haben nichts miteinander zu tun

Diese Änderung wurde bei dem gesamten Block *Gegenteils- und Synonymabfrage* durchgeführt.

Item 23 ursprünglich:	Ein*e Freund*in und ich stoppen die Zeit bei einem Wettlauf. Ich vertraue meinem Messergebnis mehr als seinem/ihrer.
Abgeändert:	Ein Kindergartenkind stoppt die Zeit bei einem Wettlauf. Ich vertraue seinem Messergebnis.
Item 24 ursprünglich:	Ein*e Wissenschaftler*in und ich stoppen die Zeit bei einem Wettlauf. Ich vertraue meinem Messergebnis mehr als seinem/ihrer.
Abgeändert:	Ein*e Physiker*in stoppt die Zeit bei einem Wettlauf. Ich vertraue seinem/ihrer Messergebnis.
Item 25 ursprünglich:	Eine Lehrperson und ich stoppen die Zeit bei einem Wettlauf. Ich vertraue meinem Messergebnis mehr als seinem/ihrer.
Abgeändert:	Eine Sportlehrer*in stoppt die Zeit bei einem Wettlauf. Ich vertraue seinem/ihrer Messergebnis.

Tabelle 7: Veränderung der Items 23, 24 und 25

Die Items 23 bis 25 werden als eigener kleiner Block betrachtet, da sie stark in Zusammenhang stehen. Bei den Interviews zeigte sich das Missverständnis, dass es sich um zwei unterschiedliche Events handelt, bei welchen die Zeit gemessen wird. Die Item Absicht liegt aber darin, den Einfluss der messenden Person auf das Messergebnis eines Events zu testen. Um dieses Missverständnis zu vermeiden, wurden die Items so umformuliert, dass in jedem der Items eine Person ein Event misst.

Schüler:in: „also...hmm...ich vertrau einfach beiden Messergebnissen“,
Interviewer:in: „Wenn sie komplett unterschiedlich wären?“
Schüler:in: „dann wäre die Lehrperson einfach schneller gelaufen und ich wäre langsamer gewesen“ (L2|2 9:40)

Interviewer:in: „Also stellst du dir das so vor, du läufst und sie misst dein Ergebnis und sie läuft und du misst ihr Ergebnis? Oder schaut ihr beide einer Person zu und messt beide das?“
Schüler:in: „Wir laufen beide gleichzeitig und haben dabei das Handy in der Hand und wenn wir an der Ziellinie sind, stoppen wir beide“ (N3 14:20)

Zudem war die Absicht bei der Itemerstellung, dass eine klare Reihenfolge zwischen den Vertrauenswürdigkeiten der Personen existiert. Diese Reihenfolge wurde jedoch nicht zwangsweise von den Schüler*innen so empfunden, wie sich zum Beispiel in folgender Aussage zeigt:

„Ein Wissenschaftler und ich stoppen die Zeit bei einem Wettlauf, ich vertraue...ist doch das gleiche [wie die Frage wo statt ‚Wissenschaftler‘
‚Freund:in‘ steht] oder nicht?“ (B1|1 5:43)

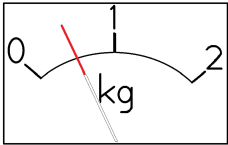
Item 26 ursprünglich:	Ich wiege beim Backen das Mehl ab. Die Waage zeigt minus 10 Gramm an. Ich halte dieses Messergebnis für vertrauenswürdig.
Abgeändert:	Ich wiege Mehl auf einer Waage ab. Die Waage zeigt folgendes an:  Ich halte dieses Messergebnis für vertrauenswürdig.
Item 27 ursprünglich:	Ich wiege eine Kilopackung (1000 Gramm) Mehl auf einer Waage ab. Die Waage zeigt 999g an. Ich halte dieses Messergebnis für vertrauenswürdig.
Abgeändert:	Eine andere Waage zeigt beim Mehl abwiegen folgendes an:  Ich halte dieses Messergebnis für vertrauenswürdig.

Tabelle 8: Veränderungen der Items 26 und 27

Auch die Items 26 und 27 sind zusammenhängend zu betrachten. In der ursprünglichen Fragestellung wurde Augenmerk auf die unterschiedlichen Zahlenwerte gelegt. Aufgrund von Gesprächen wie dem Untenstehenden wurde allerdings der Entschluss gefasst, im Folgenden den Fokus auf die Messauflösung zu legen.

Interviewer:in: „Also du hältst es nur für vertrauenswürdig weils genau ist?“

Schüler:in: „Ja.“

Interviewer:in: „Und was ist, wenn da z.B. 950g stehen würde?“

Schüler:in: „Dann würde ich dem auch vertrauen.“ (B2|1 13: 35)

Item 28 ursprünglich:	Mit einem Lineal kann ich vertrauenswürdig eine Länge von einem Zentimeter messen.
Abgeändert	Mit einem Geodreieck kann ich vertrauenswürdig eine Länge von 3cm abmessen.

Tabelle 9: Veränderung des Items 28

Bei der ursprünglichen Fragestellung kam vermehrt der Vermerk (wie in den untenstehenden Zitaten gezeigt), dass es bei Linealen manchmal große Unterschiede in der Qualität gibt. Um diese Qualitätsspanne zu verringern, wurde deswegen das „Lineal“ durch das „Geodreieck“ ersetzt. Da angenommen wird, dass bei zweiterem Spielzeug- und Werbegeschenksvarianten weniger verbreitet sind.

„Kommt drauf an woher ich das Lineal habe, weil manchmal sind Lineale sehr billig und ungenau“ (B2|1 14:24)

„Teils teils, kommt auf das Lineal an, welche Marke, wenn es irgendein Kinderlineal ist oder so, dann würde ich dem eher nicht vertrauen“ (B3 37:32)

Item 30 ursprünglich:	Eine Schulärztin misst am Montag meine Körpergröße, das Messergebnis ist 159,0 cm. Am Dienstag misst sie nochmal. Nun ist das Ergebnis 158,5 cm. Ich halte dieses Messergebnis für vertrauenswürdig.
Abgeändert	Ein*e Schularzt*in misst am Montag um 8 Uhr morgens meine Körpergröße, das Messergebnis ist 159,0 cm. Am Dienstag um 8 Uhr morgens misst sie nochmal. Nun ist das Ergebnis 157,5 cm. Ich halte diese Messergebnisse für vertrauenswürdig.

Tabelle 10: Veränderung des Items 30

Bei Item 30 merkten viele Schüler:innen an, dass die Körpergröße auch von der Tageszeit abhängt (siehe untenstehende Zitate). Da in der ursprünglichen Fragestellung nicht hervorgeht ob diese Messungen zur selben Uhrzeit stattfinden, wurde dies in der abgeänderten Variante konkretisiert.

„Ja schon, weil es ist ja unterschiedlich in der Früh, also ich hab mal gehört, dass man in der Früh größer ist als untertags“ (L3 21:40)

„Das kommt drauf an wann gemessen wurde, weil nach dem Schlafen ist die Wirbelsäule mehr gedehnt“ (E2 24:00)

„Kommt immer drauf an wie lang man schläft, weil in der Nacht ist das ja so in der Nacht wenn man wächst und in der Früh ist dann meistens größer als am Abend deswegen glaub ich eher, dass das Ergebnis in der 6. Stunde oder so gemacht haben und das 159 in der 1. Stunde“ (E4 33:17)

Item 31 ursprünglich:	Ein YouTube-Kanal dem ich vertraue, wechselt den Content-Creator (Autor). Ich vertraue dem Kanal weiterhin.
Abgeändert	Ein YouTube-Kanal dem ich vertraue, wechselt den Content-Creator (Ersteller*in). Ich vertraue dem Kanal weiterhin.

Tabelle 11: Veränderung des Items 31

Bei dem Item 31 sollte eine deutsche und eine englische Bezeichnung beibehalten werden. Allerdings gab es (wie unten zu sehen) Vermerke das „Autor“ nicht die beste Bezeichnung ist. Aus diesem Grund wurde ein, für den Kontext adäquaterer Begriff gewählt.

„[...] Content-Creator ist einfach Content-Creator, aber ein Autor schreibt Manuskripte/Bücher“ (B 2|2 18:38)

Item 9 ursprünglich:	Zwielichtigkeit und Vertrauenswürdigkeit sind für mich:
----------------------	---

Tabelle 12: Gelöschtes Items 9

Vielen Schüler:innen fehlte ein Verständnis für den Begriff Zwielichtigkeit, deswegen wurde das Item gestrichen.

„Was ist Zwielichtig“ (L1 6:10)

„Was heißt Zwielichtigkeit“ (L2 9:40)

„Frage, was heißt Zwielichtigkeit?“ (E1 6:40)

„Okay Zwielichtigkeit, versteh ich jetzt ehrlich gesagt nicht so ...irgendetwas mit Licht?“ (E3 9:00)

„Was heißt Zwielichtigkeit?“ (E4 0:46)

Item 32 ursprünglich:	Mein Großvater kann mir nicht die 9 Bundesländer Österreichs nennen. Ich vertraue ihm trotzdem.
-----------------------	---

Tabelle 13: Gelöschtes Item 32

Bei diesem Item gab es sehr unterschiedliche Ansätze, wie die Situation interpretiert wurde. Die gedachten Begründungen für das fehlende Wissen waren sehr vielfältig (zwei Beispiele sind unten angefügt) und die Antworten fielen sehr unterschiedlich aus. Aufgrund von den sehr variantenreichen Ansätzen diese Frage zu bearbeiten, wurde sie gestrichen.

„[lange Pause] hmm... ist da ein Großvater gemeint, der in Österreich geboren wurde?“ (L2|2 17:40)

“hmm... mein Großvater - vielleicht hat er Demenz oder vergisst Sachen wieder“ (E3 43:00)

4.1.7 Finaler Fragebogen

Der somit erhaltene überarbeitete dritte Fragebogen ist folgend angeführt.

Fragebogen zur Vertrauenswürdigkeit

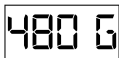
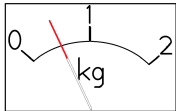
Angaben zu deiner Person

- Alter: _____
- Klasse: _____
- Geschlecht:
 - ☐ männlich
 - ☐ weiblich
 - ☐ divers
 - ☐ keine Angabe
- Ich habe bereits Physikunterricht in der Schule:
 - ☐ Ja
 - ☐ Nein

Im Folgenden findest du einige Aussagen, die du vervollständigen sollst. Wenn du eine Aussage nicht gleich verstehst, versuche trotzdem darüber nachzudenken und eine Antwort zu geben.

	sind ähnlich	sind gegenteilig	haben nichts miteinander zu tun
Unsicherheit und Vertrauenswürdigkeit...	☐	☐	☐
Richtigkeit und Vertrauenswürdigkeit...	☐	☐	☐
Glaubwürdigkeit und Vertrauenswürdigkeit ...	☐	☐	☐
Faulheit und Vertrauenswürdigkeit ...	☐	☐	☐
Verlässlichkeit und Vertrauenswürdigkeit ...	☐	☐	☐
Unehrlichkeit und Vertrauenswürdigkeit ...	☐	☐	☐
Unglaubwürdigkeit und Vertrauenswürdigkeit ...	☐	☐	☐
Langeweile und Vertrauenswürdigkeit ...	☐	☐	☐
Aufrichtigkeit und Vertrauenswürdigkeit...	☐	☐	☐
Kleine Unsicherheit beim Messen und Vertrauenswürdigkeit...	☐	☐	☐
Große Unsicherheit beim Messen und Vertrauenswürdigkeit...	☐	☐	☐

	trifft zu	trifft eher zu	teils-teils	trifft eher nicht zu	trifft nicht zu
Ich halte Informationen in meinem Schulbuch für vertrauenswürdig.	☐	☐	☐	☐	☐
Ich halte Informationen von meiner Lehrperson für vertrauenswürdig.	☐	☐	☐	☐	☐
Ich halte Informationen von meinen Eltern für vertrauenswürdig.	☐	☐	☐	☐	☐
Ich halte Informationen von einem/einer Nachhilfelehrer*in für vertrauenswürdig.	☐	☐	☐	☐	☐

Ich halte fremde Personen, die mir am Bahnsteig irgendwelche Waren anbieten wollen, für vertrauenswürdig.	☐	☐	☐	☐	☐
Ich halte Informationen von Wikipedia für vertrauenswürdig.	☐	☐	☐	☐	☐
Ich halte Informationen von YouTube-Lernvideos für vertrauenswürdig.	☐	☐	☐	☐	☐
Ich halte Informationen von unbekannten Anrufern für vertrauenswürdig.	☐	☐	☐	☐	☐
Ich halte Informationen aus dem Internet für vertrauenswürdig.	☐	☐	☐	☐	☐
Ich halte Informationen auf <i>Social Media</i> (Instagram, TikTok,...) für vertrauenswürdig.	☐	☐	☐	☐	☐
Die Antworten die mir ChatGPT (ChatBot) auf eine Frage gibt, halte ich (grundsätzlich) für vertrauenswürdig.	☐	☐	☐	☐	☐
Ich halte die Ergebnisse des Taschenrechners für vertrauenswürdig.	☐	☐	☐	☐	☐
Ich halte die Ergebnisse im Lösungsheft für vertrauenswürdig.	☐	☐	☐	☐	☐
Ein Kindergartenkind stoppt die Zeit bei einem Wettlauf. Ich vertraue seinem Messergebnis.	☐	☐	☐	☐	☐
Ein*e Physiker*in stoppt die Zeit bei einem Wettlauf. Ich vertraue seinem/ihrer Messergebnis..	☐	☐	☐	☐	☐
Eine Sportlehrer*in stoppt die Zeit bei einem Wettlauf. Ich vertraue seinem/ihrer Messergebnis.	☐	☐	☐	☐	☐
Ich wiege Mehl auf einer Waage ab. Die Waage zeigt folgendes an:	☐	☐	☐	☐	☐
					
Ich halte dieses Messergebnis für vertrauenswürdig.					
Eine andere Waage zeigt beim Mehl abwiegen folgendes an:	☐	☐	☐	☐	☐
					
Ich halte dieses Messergebnis für vertrauenswürdig.					
Mit einem Geodreieck kann ich vertrauenswürdig eine Länge von 3 cm abmessen.	☐	☐	☐	☐	☐
Ein*e Schulärzt*in misst am Montag um 8 Uhr morgens meine Körpergröße, das Messergebnis ist 159,0 cm. Am Dienstag um 8 Uhr morgens misst sie nochmal. Nun ist das Ergebnis 157,5 cm. Ich halte diese Messergebnisse für vertrauenswürdig.	☐	☐	☐	☐	☐
Draußen schneit es. Das Außenthermometer zeigt 40°C an. Ich halte dieses Messergebnis für vertrauenswürdig.	☐	☐	☐	☐	☐
Ein YouTube-Kanal dem ich vertraue, wechselt den Content-Creator (Ersteller*in). Ich vertraue dem Kanal weiterhin.	☐	☐	☐	☐	☐

4.2 Qualitätssicherung

4.2.1 Interdecoder Reliabilität Testung

Die Interdecoder Reliabilität (ICR) ist eine numerische Angabe zur Übereinstimmung von Coder:innen in Bezug darauf, wie ein Datensatz codiert wird. Die ICR wird oftmals synonym zur Interrater Reliabilität (IRR) und der Interdecoder Consistency (ICC) verwendet, auch wenn diese drei Konzepte folgendermaßen unterschieden werden können: Die IRR wird angegeben, wenn die Daten ordinal oder intervall skaliert sind, wohingegen die ICR für nominal skalierte Daten geeignet ist. Die ICC hat eine ähnliche Testabsicht wie die ICR allerdings ohne das Ziel, dass die Übereinstimmung beziffert werden kann. Die ICC ist deshalb nicht dafür geeignet, die Reliabilität anzugeben und wird stattdessen meist als ein Schritt im Prozess genutzt, um die Kategorien, in welche die Codes unterteilt werden, weiterzuentwickeln (O'Connor & Joffe, 2020).

Da in dieser Arbeit die Einschätzungen der Coder:innen nominal skaliert sind und die Reliabilität angegeben werden soll, ist die Ermittlung der ICR die richtige Vorgehensweise. Für die Erhebung der ICR gibt es mehrere Möglichkeiten. Eine der verbreitetsten Varianten ist die Angabe des Prozentsatzes von Übereinstimmungen. Allerdings wird diese Methode von Cliodhna O'Connor and Helene Joffe kritisiert, da sie nicht auf erwartete zufällige Übereinstimmungen Rücksicht nimmt. In Bezug darauf wird Krippendorff's alpha und Fleiss' κ als bessere Alternative genannt, wobei Krippendorff's alpha ausschließlich zulässig für die Ermittlung der ICR bei zwei Coder:innen ist und Fleiss' κ auch bei mehreren Coder:innen anwendbar ist (O'Connor & Joffe, 2020).

Aus diesem Grund wurde in dieser Arbeit für den Kennwert der Reliabilität das Fleiss' κ ermittelt. Die Ermittlung erfolgt folgendermaßen:

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \quad (4.1)$$

p_e ist dabei die erwartete Übereinstimmung und p_0 die beobachtete Übereinstimmung. Sie werden folgendermaßen ermittelt:

$$p_0 = \frac{1}{N \cdot n \cdot (n - 1)} \left(\left(\sum_{i=1}^N \sum_{j=1}^k n_{ij}^2 \right) - N \cdot n \right) \quad (4.2)$$

$$p_e = \sum p_j^2 \quad (4.3)$$

Dabei ist N die Anzahl der zu beurteilenden Codes, n die Anzahl der Coder:innen, k die Anzahl der Kategorien und p_j ist die statistische Wahrscheinlichkeit, dass ein Code einer bestimmten Kategorie zugewiesen wird (DATAtab Team, 2023). Das dazugehörige übliche 95%-Konfidenzintervall wird wie folgt berechnet (Cohen, 1960):

$$KI = \kappa \pm z \cdot \sigma_{\kappa} \quad (4.4)$$

Dabei ist z der für das Konfidenzniveau zu bestimmende z -Wert, welcher aus der z -Verteilungstabelle abgelesen werden muss. Im Fall eines 95%-Konfidenzintervall ist $Z = 1,96$. Weiters bezeichnet σ_{κ} die Standardabweichung von κ und wird mit der folgenden Gleichung berechnet (Cohen, 1960).

$$\sigma_{\kappa} = \sqrt{\frac{p_0(1-p_0)}{N(1-p_e)^2}} \quad (4.5)$$

Für die Interpretation vom Wert des Fleiss' κ werden Landis und Kochs Kennwerte herangezogen: (Siehe Tabelle 14 nach Landis & Koch, 1977, 163)

[Fleiss] Kappa Statistik	Maß der Überschneidung
$< 0,00$	mangelhaft
$0,00 - 0,20$	gering
$0,21 - 0,40$	mäßig
$0,41 - 0,60$	moderat
$0,61 - 0,80$	substanziell
$0,81 - 1,00$	fast perfekt

Tabelle 14: Landis und Kochs Kennwerte zur Interpretation des Fleiss' κ

Im Pre-Testing wurden 14 Lautes Denken – Interviews geführt. In jedem dieser Interviews wurden den Teilnehmenden die 34 Fragen des Fragebogens 1 gestellt. Für die ICR-Testung wurden von jeder Frage zwei Interviews zufällig ausgewählt, aus welchen die Antworten vollständig transkribiert wurden. Damit wurden 68 (pro Frage zwei) von den insgesamt 476 Antworten für die Testung herangezogen. In anderen Worten $\approx 14,3\%$ der Gesamtdaten, was im von O'Connor und Joffe genannten Richtwert von 10-25% liegt. Diese 68 Codes wurden von allen fünf Teammitgliedern kategorisiert (O'Connor & Joffe, 2020).

Im ersten Entwurf der Kategorien besteht der Versuch nicht nur zu prüfen, ob die Coder:innen sich einig sind, ob es einen Anlass gibt, die Frage zu verändern, sondern auch, ob sie bei einem Anlass das Ausmaß des Verständnisses der Befragten gleich kategorisieren. So wurden die folgenden vier Kategorien entwickelt:

- A) Die Frage oder Wörter in der Frage wird/werden nicht gemäß der Item Absicht interpretiert.
- B) Es ist unklar, ob die Frage gemäß der Item Absicht interpretiert wird — aber es gibt Anlass zur Vermutung, dass die Frage nicht gemäß der Item Absicht interpretiert werden könnte. Beziehungsweise: Es ist unklar, ob die Wörter in der Frage gemäß der Item Absicht interpretiert werden — aber es gibt Anlass zur Vermutung, dass die Wörter in der Frage nicht gemäß der Item Absicht interpretiert werden könnten.

- C) Die Frage wird gemäß der Item Absicht interpretiert — aber es gibt Anlass zur Vermutung, dass die Frage nicht gemäß der Item Absicht interpretiert werden könnte. Beziehungsweise: Die Wörter in der Frage werden gemäß der Item Absicht interpretiert — aber es gibt Anlass zur Vermutung, dass die Wörter in der Frage nicht gemäß der Item Absicht interpretiert werden könnten.
- D) Die Frage wird verstanden – und es gibt keinen Anlass zur Vermutung, dass die Frage falsch verstanden werden könnte. Beziehungsweise: Die Wörter in der Frage werden verstanden – und es gibt keinen Anlass zur Vermutung, dass die Wörter in der Frage falsch verstanden werden könnten.

Diese vier Kategorien sollen alle möglichen Fälle eindeutig abdecken. Die Kategorie D ist dabei die Einzige, welche bedeutet, dass der:die Coder:in keinen Anlass sieht, etwas an der Frage zu verändern. Die Kategorien A – C bedeuten, dass der:die Coder:in einen Anlass sieht, etwas an der Frage zu verändern. Die Kategorien unterscheiden sich lediglich darin, wie die Coder:innen das Verständnis der Befragten einschätzen. In einem ICC-Test wurde jedoch festgestellt, dass es bei der Einteilung in die Kategorien A – C eine nicht zufriedenstellende Streuung gab. Diese Streuung gab Anlass, die Kategorien neu zu definieren. Da an diesem Punkt des Prozesses das eigentliche Ziel darin bestand, herauszuarbeiten, welche Fragen verändert gehören, wurde die Einschätzung des Verständnisses der Befragten als nebensächlich angesehen. Im Zuge dessen wurden die Kategorien A-C zu einer Kategorie zusammengefasst in der Erwartung, dass eine höhere ICR erreicht wird, auch wenn diese nur noch darüber Auskunft gibt, ob Einigkeit darin besteht, dass eine Frage verändert gehört. Die neu entwickelten Kategorien wurden folgendermaßen definiert:

- A = Auffällig: Die Antwort gibt **einen** Anlass zu vermuten, dass die Frage oder Teile der Frage nicht gemäß der Item Absicht interpretiert werden oder nicht gemäß der Item Absicht interpretiert werden könnten. $\Rightarrow$ Es **gibt einen** Veränderungsbedarf bei der Frage.
- B = Unauffällig: Die Antwort gibt **keinen** Anlass zu vermuten, dass die Frage oder Teile der Frage nicht gemäß der Item Absicht interpretiert werden oder nicht gemäß der Item Absicht interpretiert werden könnten. $\Rightarrow$ Es **gibt keinen** Veränderungsbedarf bei der Frage.

Bei der Inter-coder Reliabilität Testung ergab sich für Kategorie A ein p_j -Wert von $p_A = 0,22$ und für Kategorie B der Wert $p_B = 0,77$. Somit würde mittels Gleichung 4.3 der folgende Wert für die erwartete Übereinstimmung erhalten:

$$p_e = 0,65$$

Mittels der Gleichung 4.2 und $N = 68$, $n = 5$ wurde weiters der folgende Wert für die beobachtbare Übereinstimmung ermittelt:

$$p_0 = 0,89$$

Aus diesen beiden Werten konnte der folgende Fleiss' κ Wert mit einem 95%-Konfidenzintervall mit der Gleichung 4.1 und 4.4 bzw. 4.5 berechnet werden:

$$\kappa = 0,68 \pm 0,21$$

Dieser Wert zeigt nach Tabelle 1 eine substantielle Übereinstimmung bei der Interpretation der Schüler:innen-Antworten. Nach berechnetem Konfidenzintervall kann dieser Wert jedoch von moderatem bis zu einem fast perfekten Maß der Übereinstimmung schwanken. Insgesamt zeigt dies jedoch, dass die Interpretation ausreichend reliabel ist und somit die Notwendigkeit der Änderungsmaßnahmen in substantieller Übereinstimmung beschlossen wurde.

4.2.2 Stichprobenwahl und Testing-Phase

Für die Durchführung des finalen Fragebogens wurde zuerst eine Stichprobe gewählt, welche durch die Formulierung der Forschungsfrage auf Schüler*innen der AHS Unterstufe beschränkt ist. Insgesamt wurden 273 Schüler*innen aus vier verschiedenen Schulen in Niederösterreich und Wien und insgesamt zwölf Klassen befragt. Sieben von diesen zwölf Klassen befinden sich in der 8.Schulstufe (4.Klasse), zwei in der 7.Schulstufe (3.Klasse), zwei in der 6.Schulstufe (2.Klasse) und eine in der 5.Schulstufe (1.Klasse). Diese Verteilung lässt schließen, dass es sich hierbei um eine verzerrte Stichprobe (biased sample) handelt.(Groves, 2005) Dies schränkt die Repräsentativität der Stichprobe ein, jedoch ist sie für den Rahmen einer Masterarbeit ausreichend.

Im Vorhinein wurden den Erziehungsberechtigten eine Information zur Forschungsarbeit und eine zu unterschreibende Einverständniserklärung verschickt. Beide Dokumente sind im Anhang unter D zu finden.

4.2.3 Die Faktorenanalyse - Validität

Die Daten dieser Stichprobenbefragung wurden dann einer sogenannten *Faktorenanalyse* unterzogen um die Validität zu überprüfen. Hierbei handelt es sich um ein Analyseverfahren der sozialwissenschaftlichen Forschung. Während dieses Prozesses werden Dimensionen durch die Zusammenfassung in Faktoren reduziert, wobei Daten von beobachtbaren Variablen und den Zusammenhängen zwischen ihnen erfasst werden sollen. Dadurch wird es möglich, mithilfe von beobachtbaren Größen Aussagen über nicht direkt sichtbare Faktoren zu treffen(Jordan, 2019).

Bei dem Analyseverfahren, welches zum Ansatz der R-Typ Faktorenanalyse gezählt wird, werden mithilfe des Programms SPSS Statistics Korrelationen zwischen Variablen un-

tersucht. Es entsteht eine sogenannte *Korrelationsmatrix*, bei der die Zeilen die Fälle und die Spalten die Variablen darstellen. Diese Matrix gibt einen Überblick darüber, welche Variablen in einem Zusammenhang miteinander stehen könnten. Besteht eine hohe Korrelation zwischen einzelnen Variablen, bedeutet dies, dass diese einen gemeinsamen extrahierbaren Faktor messen könnten. Bei der Faktorenextraktion werden korrelierende Variablen zu Faktoren zusammengefasst, wobei die Diagonale der Korrelationsmatrix als Ausgangspunkt für die Extraktion dient.

Des Weiteren werden in der Korrelationsmatrix die Beziehungen zwischen jeder einzelnen Variable und allen anderen Variablen geschätzt. Basierend auf diesen Kommunalitäten werden Faktoren extrahiert, um einen möglichst großen Anteil der Gesamtvarianz, bekannt als Eigenwert, abzubilden. Ein größerer Anteil der Varianzaufklärung ist mit steigenden Eigenwerten verbunden. Die Faktorladung bezeichnet die Korrelation zwischen einer Variable und einem Faktor(Schendara, 2010).

In dieser Forschungsarbeit wird die Hauptkomponentenanalyse angewandt, indem den Antwortmöglichkeiten des Fragebogens Nummern zugeordnet werden. Eine grundlegende Voraussetzung für eine sinnvolle Durchführung der Hauptkomponentenanalyse ist die Annahme, dass die Daten miteinander korrelieren. Es wird eine Linearkombination aus den numerischen Daten gebildet, die den Items zugeordnet ist, wobei diejenigen Daten verwendet werden, die die größte Varianz aufweisen, da Varianz ein Maß für die Datenstreuung ist. Diese Linearkombination wird als Projektion aus einem n -dimensionalen Raum auf eine seiner Achsen beschrieben. Anschließend werden die Informationen aus der ersten Linearkombination entfernt. Dieser Prozess kann wiederholt werden, um theoretisch so viele Hauptkomponenten zu bilden, wie Variablen vorhanden sind. Allerdings widerspricht dies dem Ziel der Dimensionsreduktion. Es gibt keine festen Regeln dafür, wie viele Linearkombinationen durchgeführt werden sollen, da dies oft vom spezifischen Kontext abhängt. Die Ladung bezeichnet die individuellen Gewichte der Variablen auf einer bestimmten Hauptkomponente. Zusätzlich kann das Koordinatensystem, in dem die Hauptkomponenten die Koordinatenachsen bilden, rotiert werden, um eine bessere Zuordnung zu den Variablen zu ermöglichen. Es gibt verschiedene Rotationsverfahren, wobei ein Beispiel die Varimax-Rotation ist, die die Orthogonalität des zugrunde liegenden Koordinatensystems beibehält. In dieser Arbeit werden verschiedene Rotationsverfahren, je nach Eindeutigkeit der Faktoren, verwendet(Gehrke, 2022).

4.2.4 Cronbachs Alpha - Reliabilität

Um die interne Konsistenz eines Fragebogens zu ermitteln, wird Cronbachs Alpha berechnet. Dieser Wert kann zwischen minus unendlich und unendlich liegen, wobei ein idealer Alpha-Wert zwischen 0,65 und 0,95 liegt.

Cronbachs Alpha ist eine statistische Methode, um die Reliabilität eines Fragebogens anzugeben.

Welche Interpretationen den unterschiedlichen Alpha-Werten zugeschrieben werden, kann der untenstehenden Tabelle 15 entnommen werden. Zusätzlich sei erwähnt, dass ein Alpha-Wert weit über 0,9 wahrscheinlich auf redundante Items hinweist.

α	Bedeutung
$> 0,9$	exzellent
$> 0,8$	gut
$> 0,7$	akzeptabel
$> 0,6$	fragwürdig
$> 0,5$	schlecht
$\leq 0,5$	inakzeptabel

Tabelle 15: Faustregel zur Interpretation der Alpha-Werte (George & Mallery, 2002)

4.2.5 Objektivität

Die Fragebögen sind so gestaltet, dass idealerweise keine externe Einflussnahme durch die Testenden erforderlich ist. Dennoch wird eine Notiz für die Forschenden verfasst, die beinhaltet, was den Schüler*innen vor dem Test mitzuteilen ist. Es wird darauf hingewiesen, dass die Schüler*innen die Fragebögen ohne Absprache mit ihren Kolleg*innen ausfüllen sollen. Wenn die Forschenden anwesend sind, wird dieser Text sinngemäß von ihnen selbst vermittelt. Werden die Fragebögen von den Lehrpersonen der Schüler*innen verteilt werden, wird ihnen der Text zusammen mit den Fragebögen übermittelt, mit der Bitte, den Inhalt sinngemäß wiederzugeben.

4.2.6 Weitere Gütekriterien

Aktionsforschung soll drei Kriterienbereiche abdecken: Erkenntnistheoretische, pragmatische und ethische Kriterien (Altrichter et al., 2018). Diese werden im Folgenden diskutiert.

Erkenntnistheoretische Kriterien

Hierbei liegt der Fokus auf dem hinzuziehen alternativen Perspektiven. Hierbei werden drei Möglichkeiten genannt: Perspektiven durch andere *Personen*, andere *Forschungsmethoden* und *Untersuchung anderer, aber ähnlicher Situationen* .(Altrichter et al., 2018)

In dieser Arbeit wurden durch das Arbeiten in der Forschungsgruppe, aber durch das Hinzuziehen von Experten alternative Perspektiven immer wieder eingebracht und eingearbeitet. Zusätzlich boten die Think-Aloud-Interviews eine alternative Forschungsmethode für eine weitere Perspektive. Somit wurden für diesen Fragebogen die erkenntnistheoreti-

schen Kriterien berücksichtigt.

Pragmatische Kriterien

Hierbei liegt der Fokus vor allem auf der Verträglichkeit mit der Praxis und die damit verbundene leichte Verständlichkeit und Handhabung (Altrichter et al., 2018).

Der Fragebogen wurde zwar mit großem Aufwand erstellt, ist jedoch für zukünftige Forschungen leicht einzusetzen. Darüber hinaus wurde der Fragebogen für die Unterstufe entwickelt und wäre somit auch in der Oberstufe einsetzbar. Die Untersuchung stört das Unterrichtsgeschehen nur für weniger als eine Unterrichtsstunde, was die Umsetzbarkeit deutlich erleichtert.

Ethische Kriterien

Bei den ethischen Kriterien soll darauf geachtet werden, dass die Forschung einerseits mit den entsprechenden pädagogischen Zielen vereinbar ist und andererseits den Grundsätzen humaner Interaktion entsprechen (Altrichter et al., 2018).

Die Schüler*innen wurden darauf hingewiesen, dass die Teilnahme am Ausfüllen der Fragebögen freiwillig ist und keine Auswirkungen auf ihre Schulnoten hat. Sie dürfen das Ausfüllen des Fragebogens jederzeit abbrechen, ohne negative Konsequenzen befürchten zu müssen. Die Daten werden anonymisiert behandelt, sodass ihre persönlichen Daten nicht in den Arbeiten erscheinen. Falls sie Fragen nicht sofort verstehen, werden sie ermutigt, darüber nachzudenken, wie die Frage gemeint sein könnte, und ihre Antworten entsprechend zu geben. Ebenfalls wird erklärt, dass die Forschenden Lehramtsstudierende sind, die zum Thema Vertrauenswürdigkeit forschen.

5 Ergebnisse

5.1 Ergebnisse des zeitverzögerten Tests

Im Folgenden werden nun die erhobenen Daten, der verzögerten Tests vorgestellt.

Insgesamt konnten 41 vollständig ausgefüllte Tests analysiert werden, von denen 12 die Unterrichtskonzeption von Hannah Loidl (HL) und 29 die von Rupert Bärenthaler-Pachner (RBP) durchgeführt haben. Vier ausgefüllte Tests konnten nicht in die Beurteilung mit einbezogen werden, da die Aufgabe 6a) nicht korrekt durchgeführt wurde. Insgesamt konnten die Schüler*innen im Mittel $10,0 \pm 2,5$ Punkte erreichen. Bei den Gruppen von RBP ergab sich ein arithmetisches Mittel von 10,5 Punkten mit einer Standardabweichung von 2,4 Punkten. Bei den Gruppen von HL ergab sich ein arithmetisches Mittel von 9,0 Punkten mit einer Standardabweichung von 2,6 Punkten. Insgesamt haben sieben der Schüler*innen von HL an dem Konzept weitergearbeitet. Bei diesen ergab sich ein Mittel von 9,6 Punkten mit einer Standardabweichung von 2,7 Punkten. Der Unterschied der beiden Mittelwerte ist nicht signifikant ($t = 0,90$). Bei den Schüler*innen von RBP hat niemand an dem Konzept weitergearbeitet.

Im Folgenden werden nun die Ergebnisse der einzelnen Aufgaben beschrieben, wobei sich nach jeder Beschreibung eine Tabelle mit den aufgeschlüsselten Ergebnissen, zu jener Aufgabe, befindet.

Aufgabe 1 konnten insgesamt 73,2% (HL: 50,0%, RBP: 82,8%) richtig beantworten. Die häufigste falsche Antwort, war hier die dritte, bei der die Stoppuhr mit $\pm 0,01$ s und die Armbanduhr mit ± 1 s in der falschen Reihenfolge waren.

		Antwort 1	Antwort 2	Antwort 3	Antwort 4
Gesamt	absolut	3	30	7	1
	relativ	7,3%	73,2%	17,1%	2,4%
Loidl	absolut	2	6	3	1
	relativ	16,7%	50,0%	25,0%	8,3%
Bärenthaler-Pachner	absolut	1	24	4	0
	relativ	3,4%	82,8%	13,8%	0,0%

Tabelle 16: Aufgeschlüsselte Ergebnisse: Aufgabe 1 (richtige Antwort hervorgehoben)

Bei Aufgabe 2 entschieden sich insgesamt 29,3% (HL: 16,7% und RBP: 34,5%) für die richtige Antwort. Hier wurde Antwort 1 von den Schüler*innen von Loidl am häufigsten gewählt mit 58,3%.

		Antwort 1	Antwort 2	Antwort 3	Antwort 4
Gesamt	absolut	13	11	12	5
	relativ	31,7%	26,8%	29,3%	12,2%
Loidl	absolut	7	3	2	0
	relativ	58,3%	25,0%	16,7%	0,0%
Bärenthaler-Pachner	absolut	6	8	10	5
	relativ	20,7%	27,6%	34,5%	17,2%

Tabelle 17: Aufgeschlüsselte Ergebnisse: Aufgabe 2 (richtige Antwort hervorgehoben)

Aufgabe 3 konnten insgesamt 56,1% (HL: 41,7% und RBP: 62,1%) zur Gänze richtig beantworten. Dabei wurden die drei richtigen Antworten insgesamt von jeweils mehr als 80% genannt. Die Falschantworten wurden jeweils von weniger als 20% genannt.

		Antwort 1	Antwort 2	Antwort 3	Antwort 4	Antwort 5
Gesamt	absolut	35	34	3	34	6
	relativ	85,4%	82,9%	7,3%	82,9%	14,6%
Loidl	absolut	10	8	1	9	2
	relativ	83,3%	66,7%	8,3%	75,0%	16,7%
Bärenthaler-Pachner	absolut	25	26	2	25	4
	relativ	86,2%	89,7%	6,9%	86,2%	13,8%

Tabelle 18: Aufgeschlüsselte Ergebnisse: Aufgabe 3 (richtige Antworten hervorgehoben)

Insgesamt 41,5% (HL: 41,7%, RBP: 41,4 %) konnten Aufgabe 4 korrekt beantworten. Die beiden falschen Aussagen wurden dabei von weniger als 20% ausgewählt.

		Antwort 1	Antwort 2	Antwort 3	Antwort 4	Antwort 5
Gesamt	absolut	32	6	7	30	27
	relativ	78,0%	14,6%	17,1%	73,1%	65,9%
Loidl	absolut	8	2	2	8	8
	relativ	66,7%	16,7%	16,7%	66,7%	66,7%
Bärenthaler-Pachner	absolut	24	4	5	22	19
	relativ	82,8%	13,8%	17,2%	75,9%	65,5%

Tabelle 19: Aufgeschlüsselte Ergebnisse: Aufgabe 4 (richtige Antworten hervorgehoben)

Aufgabe 5 konnten insgesamt 39,0% (HL: 25,0%, 44,8%) richtig beantworten. Bei den Schüler*innen von Loidl wurde die falsche Antwort 3 von 41,7% gewählt, also öfter als das

Eineinhalbfache der richtige Antwort.

		Antwort 1	Antwort 2	Antwort 3	Antwort 4	Antwort 5
Gesamt	absolut	13	16	7	2	2
	relativ	31,7%	39,0%	17,1%	4,9%	4,9%
Loidl	absolut	1	3	5	2	1
	relativ	8,3%	25,0%	41,7%	16,7%	8,3%
Bärenthaler-Pachner	absolut	12	13	2	0	1
	relativ	41,4%	44,8%	6,9%	0,0%	3,4%

Tabelle 20: Aufgeschlüsselte Ergebnisse: Aufgabe 5 (richtige Antwort hervorgehoben)

Der Teil von Aufgabe 6a, bei der die Schüler*innen die Vertrauenswürdigkeit einschätzen sollten, wurde von insgesamt 90,2% (HL: 83,3%, RBP: 93,1%) korrekt beantwortet.

		gar nicht vertrauenswürdig	vertrauenswürdig	sehr vertrauenswürdig
Gesamt	absolut	37	3	1
	relativ	90,2%	7,3%	2,4%
Loidl	absolut	10	1	1
	relativ	83,3%	8,3%	8,3%
Bärenthaler-Pachner	absolut	27	2	0
	relativ	93,1%	6,9%	0,0%

Tabelle 21: Aufgeschlüsselte Ergebnisse: Aufgabe 6a Vertrauenswürdigkeit einschätzen (richtige Antworten hervorgehoben)

Der Teil der Aufgabe 6a, bei der Verbesserungsvorschläge gemacht werden sollten, wurde von 95,1% (HL: 100%, RBP: 93,1) richtig bearbeitet. Am Häufigsten wurde hierbei ein Verbesserungsvorschlag aus der Kategorie K-2.3 mit über 90% angegeben.

		K-1.1	K-1.4	K-2.1	K-2.2	K-2.3	K-3.1	K-3.2
Gesamt	absolut	26	24	1	2	38	2	0
	relativ	63,4%	58,5%	2,4%	4,9%	92,7%	4,9%	0,0%
Loidl	absolut	8	8	0	0	12	0	0
	relativ	66,7%	66,7%	0,0%	0,0%	100,0%	0%	0%
Bärenthaler-Pachner	absolut	18	16	1	2	26	2	0
	relativ	62,1%	55,2%	3,4%	6,9%	89,7%	6,9%	0,0%

Tabelle 22: Aufgeschlüsselte Ergebnisse: Aufgabe 6a offene Aufgabe (richtige Antworten hervorgehoben)

Beim ersten Teil der Aufgabe 6b, der Vergleich der Vertrauenswürdigkeit, wurde von 80,5% (HL: 83,3%, RBP: 79,3%) korrekt bearbeitet, wobei ca. zwei Drittel davon eine korrekte Begründung liefern konnte.

		Antwort 1	Antwort 2	Antwort 3	Richtige Begründung	Falsche Begründung
Gesamt	absolut	2	6	33	28	4
	relativ	4,9%	14,6%	80,5%	68,3%	9,8%
Loidl	absolut	1	1	10	8	4
	relativ	8,3%	8,3%	83,3%	66,7%	33,3%
Bärenthaler-Pachner	absolut	1	5	23	20	9
	relativ	3,4%	17,2%	79,3%	69,0%	31,0%

Tabelle 23: Aufgeschlüsselte Ergebnisse: Aufgabe 6b Vergleich von Vertrauenswürdigkeit (richtige Antworten hervorgehoben)

Der letzte Teil von Aufgabe 6b wurde von 90,2% (HL: 83,3%, RBP: 93,1%) korrekt bearbeitet. Am Häufigsten wurde hier ein Verbesserungsvorschlag der Kategorie K-1.1 genannt.

		K-1.1	K-1.2	K-2.1	K-2.2	K-2.3	K-2.4	K-3.1
Gesamt	absolut	17	0	8	1	2	6	4
	relativ	41,5%	0,0%	19,5%	2,4%	4,9%	14,6%	9,8%
Loidl	absolut	4	0	3	0	0	3	2
	relativ	33,3%	0,0%	25,0%	0,0%	0,0%	25,0%	16,7%
Bärenthaler-Pachner	absolut	13	0	5	1	2	3	2
	relativ	44,8%	0,0%	17,2%	3,4%	6,9%	10,3%	6,9%

Tabelle 24: Aufgeschlüsselte Ergebnisse: Aufgabe 6b offene Aufgabe (richtige Antworten hervorgehoben)

Beim Vergleich der aktuellen Ergebnisse mit denen von damals, lässt sich definitiv ein Abfall erkennen. So konnte damals die erste Aufgabe bei Bärenthaler-Pachner von 100% richtig beantwortet werden.

Aufgabe 2 zeigt von allen Aufgaben den stärksten Abfall, von 95,8% auf 29,3%.

Aufgabe 3 zeigt eine Verbesserung von 42,9% auf 56,1%.

Aufgabe 4 wurde stark abgewandelt. Verglichen mit der Aufgabe von Bärenthaler-Pachner, die dasselbe Lernziel hatte, ist hier das Ergebnis von 85,7% auf 41,5% gesunken.

Da Aufgabe 5 eine gänzlich neue Aufgabe war, gibt es hierzu keine Vergleichswerte.

Aufgabe 6 lässt sich mit den Ergebnissen von Loidl gut vergleichen, da hier beinahe keine Änderungen an der Angabe vorgenommen wurden. Die Vertrauenswürdigkeit wurde

insgesamt korrekter beurteilt, als damals. So stieg hier das Ergebnis von 76,0% auf 90,2%. Zu den Kategorien gibt es von Loidl keine konkreten Werte. Einzig die Kategorie K-2.2 wurde damals von 50% genannt und ist bei diesem Test nur mehr von 4,9% vertreten. Besonders bei 6b hat sich eine positive Entwicklung getan. Damals konnten bei Loidl auch ca. 80% die richtige Antwort ankreuzen, jedoch konnte damals keine einzige Person eine richtige Begründung nennen. Bei den Verbesserungsvorschlägen gibt es erneut keine konkreten Werte von Loidl.

5.2 Ergebnisse des Fragebogens zur Vertrauenswürdigkeit

Im Folgenden werden nun die Ergebnisse der vorliegenden Studie vorgestellt. Hierbei wird zuerst die deskriptive Statistik der Synonym- und Gegenteilabfrage vorgestellt. Danach folgt die Analyse der Konstrukte, welche die Auswertung der Validität und der Reliabilität beinhaltet. Im Zuge dessen wird ebenfalls der Ausschluss von zwei Items begründet. Anschließend werden die Ergebnisse der metrischen Daten gezeigt und analysiert. Zuletzt werden noch separat die Ergebnisse der Items, die im Zusammenhang mit *Messungen* stehen, veranschaulicht.

5.2.1 Synonym- und Gegenteilabfrage

Zunächst werden die nominalen Daten näher betrachtet. Es wird die absolute Häufigkeit betrachtet, mit der die Möglichkeiten „sind ähnlich“, „sind gegenteilig“, „haben nichts miteinander zu tun“ angekreuzt wurden. Zusätzlich gibt es ungültige Antworten, die dann eintreten, wenn gar keine oder mehr als eine Antwortmöglichkeit gewählt wurde. Im Folgenden werden die Kategorien aufgrund der leichteren Lesbarkeit folgendermaßen abgekürzt: „sind ähnlich“ wird zu „**synonym**“, „sind gegenteilig“ wird zu „**gegenteilig**“ und „haben nichts miteinander zu tun“ wird zu „**unabhängig**“.

Ergebnisse In Abb.1 ist ein Überblick über die Ergebnisse aller Attribute gegeben. Die Reihenfolge entspricht hier nicht der auf dem Fragebogen, sondern wurde von der Autorin angepasst um auf gewählte Kategorien hinzuweisen. Im Folgenden werden allgemein die Ergebnisse dargestellt. Die genauere Betrachtung dieser folgt anschließend in der Detailanalyse.

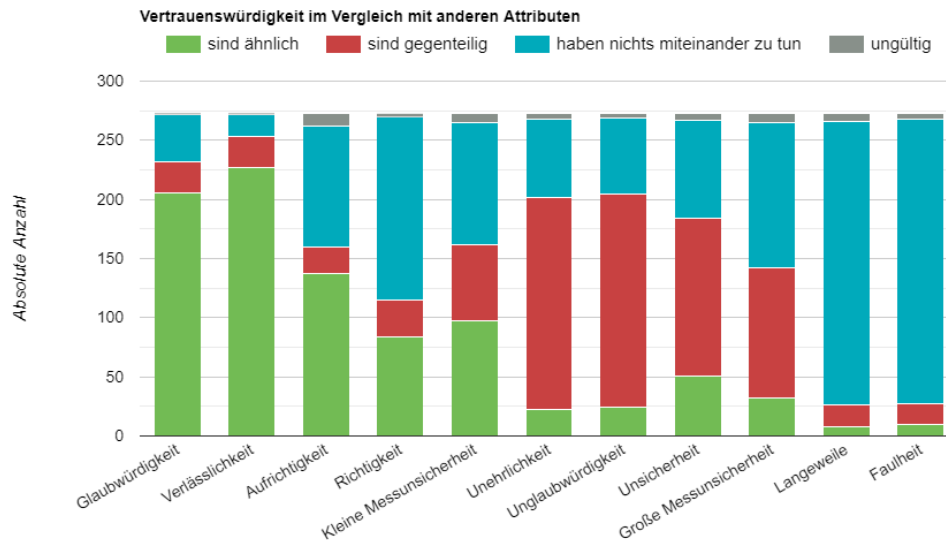


Abbildung 1: Vertrauenswürdigkeit im Vergleich mit anderen Attributen

Bei den ersten fünf Attributen handelt es sich um Attribute, welche im Vorhinein als Synonyme festgelegt wurden. Die Begriffe „Glaubwürdigkeit“ und „Verlässlichkeit“ spiegeln diese Annahme auch eindeutig wider. Bei dem Begriff der Aufrichtigkeit entschieden sich knapp über 50% für die Antwortmöglichkeit „sind ähnlich“. Bei den Attributen „Richtigkeit“ und „Kleine Messunsicherheit“ waren es jeweils weniger als ein Drittel.

Die folgenden vier Attribute stellen die Gegenteile dar. Hier gibt es wieder bei den Begriffen „Unehrllichkeit“ und „Unglaubwürdigkeit“ eindeutige Ergebnisse. Bei „Unsicherheit“ entschieden sich knapp unter 50% für die Option „gegenteilig“. Bei der großen Messunsicherheit waren es knapp 40%.

Die beiden unabhängigen Attribute „Langeweile“ und „Faulheit“ wurden auch eindeutig als solche erkannt.

5.2.2 Analyse der Konstrukte

Konstruktvalidität (Faktorenanalyse) Für die Faktorenanalyse wurden die Variablen in die beiden Konstruktsseiten *menschlich* und *künstlich* eingeteilt und jeweils eine Analyse durchgeführt. Sollten bestimmte Variablen auf gar keinen Faktor projizieren, dann werden diese Variablen ausgeschlossen und die Analyse wird wiederholt. Um ein besseres Ergebnis zu erhalten, wird eine Rotation eingeführt. Die Rotation zielt darauf ab, die Faktorladungen zu optimieren, indem sie die Ausprägung der Ladungen auf den einzelnen Variablen maximiert oder minimiert, je nach gewählter Methode. Dies führt dazu, dass jede Variable mit möglichst wenigen Faktoren stark korreliert und somit die Interpretation erleichtert wird. Die erste Faktorenanalyse mit allen Variablen, die sich auf das „künstlich“ beziehen, ist in der Tabelle 25 zu sehen.

Die Variable „Schnee“ bildet auf keine der beiden Faktoren ab. Auch weitere Untersuchungen mit unterschiedlicher Anzahl von Faktoren lieferte keine dementsprechende Abbildung. Daher wird die Analyse mit Ausschluss dieser Variable wiederholt. Die Ergebnisse sind in Tabelle 26 zu sehen.

Hier kann nun betrachtet werden, auf welchen Faktor die jeweilige Variable am stärksten projiziert. Die Interpretation davon ist im Kapitel 6.2 zu finden. Die farbliche Markierung in der Tabelle 26 entspricht der jeweiligen Zugehörigkeit zu den beiden Faktoren. Orange markiert hierbei die Komponente 1, die im Folgenden als „Dingkomponente 1“ bezeichnet wird, während Cyan die Komponente 2 kennzeichnet, die als „Dingkomponente 2“ bezeichnet wird.

Strukturmatrix		
	Komponente	
	1	2
Schulbuch	,141	,509
Wikipedia	,471	,243
YouTube	,677	,145
Internet	,702	,135
Social media	,741	-,117
ChatGPT	,527	,152
Taschenrechner	,427	,635
Lösungsheft	,296	,590
Mehl digital	,056	,664
Mehl mechanisch	,017	,476
Geodreieck	,206	,549
Schnee	,028	-,148

Tabelle 25: Faktoranalyse für alle Variablen „künstlich“. Extraktionsmethode: Hauptkomponentenanalyse. Rotationsmethode: Oblimin mit Kaiser-Normalisierung.

Strukturmatrix		
	Komponente	
	1	2
Schulbuch	,136	,502
Wikipedia	,468	,275
YouTube	,678	,194
Internet	,703	,185
Social media	,740	-,047
ChatGPT	,532	,178
Taschenrechner	,414	,662
Lösungsheft	,285	,605
Mehl digital	,033	,675
Mehl mechanisch	,007	,470
Geodreieck	,197	,553

Tabelle 26: Faktoranalyse für „künstlich“ ohne die Variable „Schnee“. Extraktionsmethode: Hauptkomponentenanalyse. Rotationsmethode: Oblimin mit Kaiser-Normalisierung.

Dasselbe Vorgehen wird mit den „menschlichen“ Variablen wiederholt. Die Tabelle 27 zeigt die Analyse mit allen Variablen.

Mustermatrix			
	Komponente		
	1	2	3
Lehrperson	,614	-,403	-,116
Eltern	,302	-,455	,228
Nachhilfelehrer	,720	-,054	-,288
Bahnsteig	,077	,745	,054
Anrufer	-,038	,702	-,166
Kindergarten	,100	,019	,676
Physiker	,691	,064	,110
Sportlehrer	,632	,215	,197
Schulärztin	-1,89	-,134	,770
YouTube-Kanal	,199	,423	,321

Tabelle 27: Faktorenanalyse für alle Variablen „menschlich“. Extraktionsmethode: Hauptkomponentenanalyse. Rotationsmethode: Oblimin mit Kaiser-Normalisierung. Rotation ist in 11 Iterationen konvergiert.

Hierbei bilden „Kindergarten“ und „Schulärztin“ als einzige stark auf die dritte Komponente ab. Dies macht jedoch aus Sicht der Interpretation keinen Sinn, denn mit diesen beiden Items sollen grundsätzlich unterschiedliche Dinge überprüft werden. Das Item der Schulärztin soll überprüfen, ob die Schüler:innen Messfehler erkennen können und sich bewusst sind, dass eines der beiden oder vielleicht sogar beide Ergebnisse nicht stimmen können. Das Item des Kindergartenkindes hingegen soll überprüfen, ob sie einem Kindergartenkind zutrauen würden, ein gutes Messergebnis beim Wettlauf zu erzielen. Es handelt sich also um unterschiedliche Fragen, weshalb es keinen Sinn macht, beide in einer Gruppe zusammenzufassen. Daher wurde entschieden, dass die Analyse ohne die Variable „Schulärztin“ wiederholt wird. Dies ist in Tabelle 28 zu sehen.

Rotierte Komponentenmatrix			
	Komponente		
	1	2	3
Lehrperson	,755	,124	-,186
Eltern	,625	,012	-,141
Nachhilfelehrer	,762	,075	,167
Bahnsteig	-,187	,103	,729
Anrufer	-,169	-,177	,709
Kindergarten	-,125	,622	-,003
Physiker	,284	,706	-,004
Sportlehrer	,131	,788	,101
YouTube-Kanal	,165	,146	,620

Tabelle 28: Faktorenanalyse für „menschlich“, ohne „Schulärztin“. Extraktionsmethode: Hauptkomponentenanalyse. Rotationsmethode: Varimax mit Kaiser-Normalisierung. Rotation ist in 5 Iterationen konvergiert.

Nun kann eine Zuordnung zu den Faktoren stattfinden, die sich auch sinnvoll interpretieren lässt. Es entstehen dadurch drei Komponenten, die mit den Farben grün, blau und rot markiert sind. Komponente 1, welche grün ist, wird im Folgenden als „Personenkomponente 1“ bezeichnet, Komponente 2 (blau) als „Personenkomponente 2“ und Komponente 3 (rot) als „Personenkomponente 3“.

Reliabilität Grundsätzlich gilt, dass ein Alpha-Wert zwischen 0.65 und 0.95 anzustreben ist. Daher wird zunächst eine Analyse mit allen 22 metrischen Variablen durchgeführt. Tabelle 29 zeigt die Fallverarbeitung und Tabelle 30 den Alphawert.

Zusammenfassung der Fallverarbeitung

		N	%
Fälle	Gültig	230	84,2
	Ausgeschlossen	43	15,8
	Gesamt	273	100,0

Tabelle 29: Fallverarbeitung aller Items. Listenweise Löschung auf der Grundlage aller Variablen in der Prozedur

Reliabilitätsstatistiken

Cronbachs Alpha	Anzahl der Items
,716	22

Tabelle 30: Alpha-Wert aller Items

Der Alphawert von 0,716 fällt nach Tabelle 15 in die Kategorie *akzeptabel*. Da jedoch bei der Faktorenanalyse die beiden Items „Schulärztin“ und „Schnee“ ausgeschlossen wurden, aufgrund fehlender Übereinstimmung mit einem Faktor, werden auch diese Items aus der Alpha-Wert-Analyse entfernt. Tabelle 31 zeigt die dazugehörigen Ergebnisse.

Reliabilitätsstatistiken

Cronbachs Alpha	Anzahl der Items
,738	20

Tabelle 31: Alpha-Wert aller Items

Um den Alpha-Wert und die damit zusammenhängende Reliabilität weiter zu verbessern, kann zusätzlich noch eine Item-Skala-Statistik erstellt werden. Diese berechnet unter anderem den jeweiligen Alpha-Wert, wenn ein bestimmtes Item weggelassen wird. Die beiden Items, die den Alpha-Wert jedoch erhöhen würden (Wie in Tab.32 zu sehen), sind jedoch gut in der Faktorenanalyse integriert, weshalb es in diesem Fall nicht sinnvoll wäre, diese zu entfernen.

Item-Skala-Statistik	
	Cronbachs Alpha, wenn Item weggelassen
Schulbuch	,721
Lehrperson	,723
Eltern	,733
Nachhilfelehrer	,723
Bahnsteig	,739
Wikipedia	,724
YouTube	,723
Anrufer	,744
Internet	,721
Social media	,725
ChatGPT	,730
Taschenrechner	,716
YouTube-Kanal	,728
Lösungsheft	,720
Kindergarten	,743
Physiker	,722
Sportlehrer	,715
Mehl digital	,731
Mehl mechanisch	,736
Geodreieck	,731

Tabelle 32: Item-Skala-Statistik, optische Trennung von „Quellen“ und „Messungen“

Ausschluss von Items Aufgrund der Ergebnisse der Faktorenanalyse und des Alpha-Werts werden für alle folgenden Analysen und auch Interpretationen die beiden Items

„Schnee“ und „Schulärztin“ ausgeschlossen. Die Bedeutung dieses Ausschlusses wird dann noch eingehend in der Diskussion behandelt.

5.2.3 Vertrauenswürdigkeit von Quellen

Die ersten 14 Items in Tabelle 32 beschäftigen sich alle mit einer Art „Informationsquelle“. Die Schüler*innen sollten hier einschätzen, für wie vertrauenswürdig sie Informationen von unterschiedlichen Quellen halten. Dabei sollen sowohl Personen als auch Dinge eingeschätzt werden.

Ergebnisse Die folgende Abbildung 2 zeigt die deskriptive Statistik der Items, die sich mit Informationsquellen beschäftigen. Der Antwortmöglichkeit „trifft zu“ wurde dabei die Zahl 1 zugeordnet, „trifft eher zu“ die Zahl 2 und so weiter. Die vertrauenswürdigste Quelle ist hierbei der Taschenrechner mit einem Mittelwert von $1,33 \pm 0,70$, die am wenigsten vertrauenswürdigste Quelle ist der unbekannte Anrufer mit dem Mittelwert $4,78 \pm 0,72$, dicht gefolgt von der fremden Person am Bahnsteig mit $4,68 \pm 0,72$.

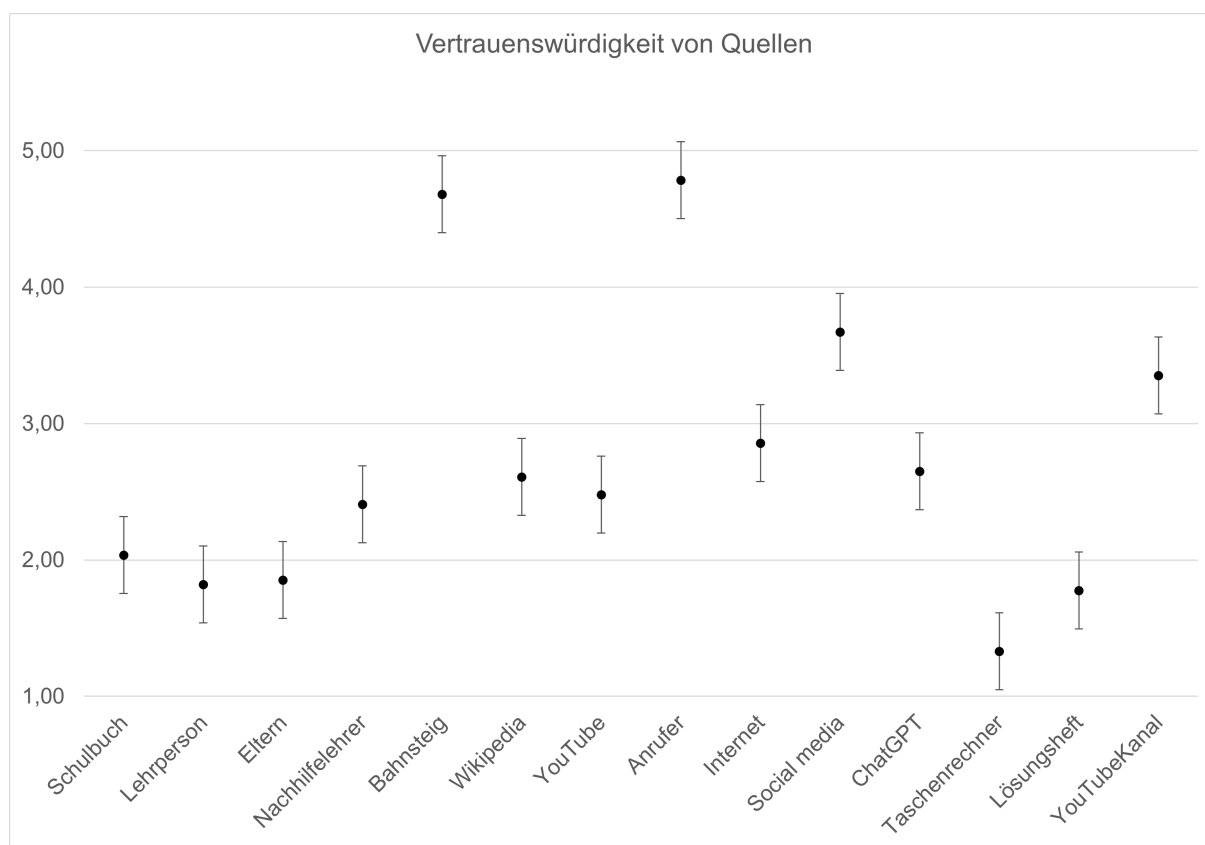


Abbildung 2: Deskriptive Statistik der Items zu „Quellen“, „trifft zu“ = 1, . . . , „trifft nicht zu“ = 5

5.2.4 Vertrauenswürdigkeit von Messungen

Die sechs Items der zweiten Gruppe in Tabelle 32 beschäftigen sich mit Messungen auf unterschiedliche Weisen. Die Schüler*innen sollten hier einschätzen für wie vertrauenswürdig

sie Messergebnisse in unterschiedlichen Situationen halten.

Ergebnisse Abbildung 3 zeigt die deskriptive Statistik der Items die sich mit Messungen auseinandersetzen. Wie auch bei den Informationsquellen wurde der Antwortmöglichkeit „trifft zu“ wurde dabei die Zahl 1 zugeordnet, „trifft eher zu“ die Zahl 2 und so weiter. Das Geodreieck bildet hierbei die vertrauenswürdigste Quelle mit einem Mittelwert von $1,49 \pm 0,77$, die am wenigsten vertrauenswürdigste Quelle ist mit großem Abstand das messende Kindergartenkind mit einem Mittelwert von $3,55 \pm 1,20$.

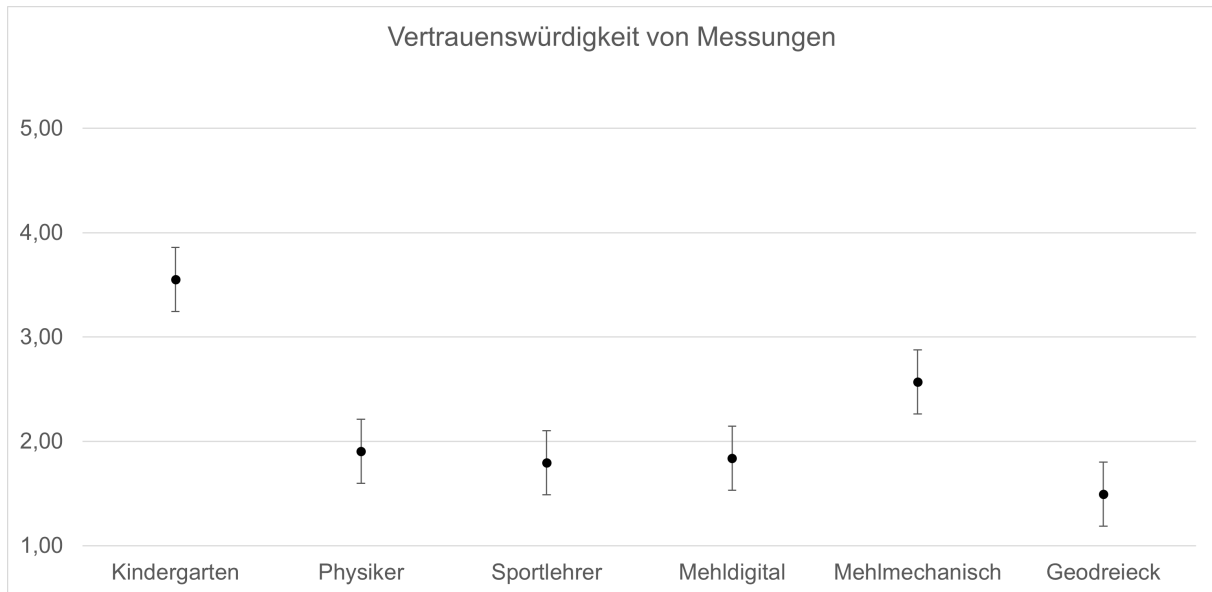


Abbildung 3: Deskriptive Statistik der Items zu „Messungen“ (ohne „Schulärztin“ und „Schnee“)

6 Diskussion

6.1 Zeitverzögerter Test

Die Ergebnisse des zeitverzögerten Tests erscheinen auf den ersten Blick eher ernüchternd. Wenn die Ergebnisse auf einen typischen Notenschlüssel übertragen würden, läge der Durchschnitt der Punkte gerade noch im Bereich „Befriedigend“. Positiv zu vermerken ist jedoch, dass nur fünf der Tests insgesamt negativ bewertet wurden. Die Tatsache, dass die ersten vier Aufgaben, nur leicht verändert, von Bärenthaler-Pachner übernommen wurden, spiegelt sich auch in den Ergebnissen wider. Schüler*innen, die diese Beispiele bereits dreiviertel Jahr zuvor bearbeitet hatten, erzielten erwartungsgemäß bessere Ergebnisse, wie in Kapitel 5.1 ersichtlich ist.

Bei Aufgabe 6, die nahezu identisch von Loidl übernommen wurde, trat dieser Effekt nicht auf. Zwar erzielten die Schüler*innen von Loidl bei den Verbesserungsvorschlägen zu Teilaufgabe 6a bessere Ergebnisse, jedoch erreichten die Schüler*innen von Bärenthaler-Pachner im Durchschnitt bei den anderen Teilaufgaben mehr Punkte. Dies ist vermutlich auf den generellen Altersunterschied der Schüler*innen zurückzuführen.

Die beiden Aufgaben, die das Einschätzen von Vertrauenswürdigkeit verlangten (Aufgabe 1 und Aufgabe 6a), zeigten insgesamt erfreuliche Ergebnisse. Diese beiden Aufgaben bilden auch mehr oder weniger das Herzstück der gesamten Thematik. Denn solange Schüler*innen die Vertrauenswürdigkeit, also die Messunsicherheit einer Messung, nicht einschätzen und beurteilen können, ist auch jedes Wissen über die Unsicherheitstypen und die korrekte Schreibweise hinfällig. So gesehen, lieferten die zeitverzögerten Test dahingehend auch gewünschte Ergebnisse.

Die Ergebnisse von Aufgabe 2 entsprachen auch den Erwartungen. Da es sich hierbei um eine bestimmte Schreibweise handelt, also um ein zu lernendes Wissen, war voraus zu sehen, dass dieses bei den meisten Schüler*innen nicht mehr vorhanden war. Interessant ist hier, dass am häufigsten Antwort 1 gewählt wurde. Eines der Ergebnisse bei Antwort 1 war $v = 130 \text{ km/h}$, welches ein Ergebnis ist, welches den Schüler*innen aus diversen Rechenbeispielen sowohl aus dem Physik als auch aus dem Mathematikunterricht bereits vertraut ist. Dies könnte eine Erklärung sein, warum diese Antwort auch am Häufigsten gewählt wurde.

Was bei den beiden Aufgaben 3 und 4 wirklich positiv hervorzuheben ist, ist die geringe Menge an falschen Antworten. Zwar konnten die beiden Aufgaben nur von rund 40% komplett richtig gelöst werden, jedoch kreuzten insgesamt weniger als 30% jeweils falsche Antworten an. Es wurden auch jeweils nur von einer Person beide falsche Antworten

gekreuzt. Dies lässt darauf schließen, dass die Schüler*innen eine grundlegende Idee haben, wie Messunsicherheiten einerseits entstehen und auch, wie sich Typ-A und Typ-B unterscheiden.

Die Ergebnisse von Aufgabe 5 sind schnell zu erklären. Obwohl Antwort 2 zwar die höhere Auflösung hatte, befanden sich bei Antwort 1 mehr Kommastellen, aufgrund der unterschiedlichen Einheiten. Die Anzahl der Kommastellen lassen Schüler*innen glauben, dass das Ergebnis *besser* ist, wie auch schon Kok (Kok et al., 2019) in ihrem Artikel zeigte.

Sowohl die Verbesserungsvorschläge bei 6a als auch bei 6b sind äußerst positive Ergebnisse. Die Schüler*innen verlangen vor allem genauere Messgeräte und mehrere Messwerte um zu einem vertrauenswürdigerem Ergebnis zu kommen. Genau hier liegt auch das Herzstück der Unterrichtskonzeptionen. Die Tatsache, dass sich zwei Drittel und mehr genau diese Aspekte aus den Konzeptionen mitgenommen haben, ist genau jenes Ziel, welches damit verfolgt wurde

Der extreme Anstieg der richtigen Begründungen bei 6b von 0% auf knapp 70% ist hier wohl auf die Art der Korrektur zurückzuführen. Leider ist von Loidl nicht bekannt, welche Aussagen genau als richtig oder falsch gewertet wurden. Bei diesem Test wurden alle Begründungen, welche sinngemäß "zu wenige Messungen" oder "zu wenige Informationen (über die Messunsicherheit" wiedergaben, als richtig gewertet.

Insgesamt sind die Ergebnisse des zeitverzögerten Test äußerst erfreulich, da die beiden Kern-Key-Ideas bei einem Großteil der Schüler*innen nach wie vor vorhanden sind. Die Tatsache, dass die wenigsten von ihnen an diesem Konzept weitergearbeitet haben, lässt auch die Aussagen treffen, dass diese Key-Ideas noch von den damals durchgeführten Konzeptionen übrig geblieben sind. Der Leistungsunterschied zwischen den beiden Konzeptionsgruppen ist wohl hauptsächlich auf den Altersunterschied und dem Unterschied des Fokus bei den Unterrichtskonzeptionen zurück zu führen.

6.1.1 Limitationen des Tests

Zwei ganz gravierende Aspekte schränken die Interpretation des Tests ein. Zum Einen wurde der Test nur von 41 Schüler*innen durchgeführt, wobei auch nur zwölf Schüler*innen von Loidl dran teilnahmen. Dies schränkt Aussagen über Unterschiede zwischen den Gruppen stark ein, da diese in den meisten Fällen nicht statistisch signifikant, aufgrund der Stichprobengröße, sind. Zum Anderen fehlt hier eine Art Kontrollgruppe um wirklich Aussagen über die positive Auswirkung der Unterrichtskonzepte zu treffen. Hier wäre es interessant zu wissen, wie Schüler*innen derselben Schulstufe, vielleicht sogar selben Klasse, abschneiden, die diese Unterrichtskonzeption niemals erfahren haben.

Zusätzlich muss noch darauf hingewiesen werden, dass bei den Schüler*innen von Loidl eine Zeitspanne von fast zwei Jahren zwischen der Durchführung der Konzeption und dem Test lag, wohingegen bei den Schüler*innen von Bärenthaler-Pachner nur ein dreiviertel Jahr dazwischen gelegen ist. Auch wirkt sich die Tatsache auf die Ergebnisse der Tests aus, dass eine Gruppe einen Teil der Aufgaben bereits aus der Unterrichtskonzeption kennt und ein anderer Teil nicht.

6.2 Fragebogen zur Vertrauenswürdigkeit

Die Ergebnisse des Fragebogens zeigen, dass das Verständnis für Vertrauenswürdigkeit bei den Schüler*innen der SEKI bereits die Grundlagen besitzt, die für den Unterricht gebraucht werden.

Die Vertrauenswürdigkeit von Quellen wurde von den Schüler*innen den Erwartungen entsprechend eingeschätzt. Dies zeigt ein entsprechendes Alltagsverständnis des Begriffs. Vor allem die Tatsache, dass der Taschenrechner die vertrauenswürdigste Quelle darstellt, zeigt, dass die Schüler*innen den menschlichen Fehler in ihre Abschätzungen miteinbeziehen. Dass die beiden unbekannten Quellen auch die am wenigstens vertrauenswürdigsten sind, zeigt außerdem, dass die Schüler*innen wenig Vertrauen zeigen, wenn ihnen entsprechende Informationen fehlen. Bereits diese Ergebnisse bestärken das untersuchte Brückenkonzept.

Die Abfrage der Attribute bestätigt dieses Verständnis erneut. Auch hier wurden, bis auf eine Ausnahme, die Antworten im Mittel den Erwartungen entsprechend gegeben. Besonders hervorzuheben sind hier die beiden Attribute „Große Unsicherheit beim Messen“ und „Kleine Unsicherheit beim Messen“. Zwar leicht abgeändert aber im Kern trotzdem gleich, decken diese beiden nämlich das Ziel des Brückenkonzepts ab. Hier wurde ohne jegliches Vorwissen der Begriff „Vertrauenswürdigkeit“ insgesamt richtig zugeordnet. Hierbei handelt es sich um ein Ergebnis, welches zwar erhofft wurde, jedoch nicht mit Sicherheit davon ausgegangen wurde.

Zuletzt jedoch überzeugten die Items mit den Vertrauenswürdigkeiten von Messungen. Diese zeigten, dass die Schüler*innen dieser doch sehr personenbehafteten Begriff (wie in den Brainstormings gezeigt) richtig den Messungen zuordnet konnten. Auch ein grundlegendes Verständnis von Typ-B-Messunsicherheiten wurde bei den beiden Mehl-Items gezeigt. Dass das Geodreieck das vertrauenswürdigste Messinstrument darstellte, lässt sich dadurch erklären, dass die Schüler*innen speziell in der SEKI im Mathematikunterricht regelmäßig mit diesem Instrument arbeiten und es dementsprechend auch besser einschätzen können als beispielsweise eine Waage.

Besonders interessant war die Faktorenanalyse der Personen. Hier sind drei Faktoren

entstanden: Personenkomponente 1, 2 und 3. Vergleicht man nun die Bestandteile der Komponenten mit den Mittelwerten der Daten, so kann Komponente 1 als „vertrauenswürdige Personen“ und Komponente 3 als „nicht vertrauenswürdige Personen“ beschrieben werden. Äußerst spannend ist hier jedoch die Personenkomponente 2. Diese beinhaltet die drei messende Personen Kindergartenkind, Physiker und Sportlehrer. Die Schüler*innen dürften also zwischen Personen als Informationsträger und Personen als „Messgeräte“ unterscheiden. Dies lässt darauf schließen, dass die Dimension nicht nur in Menschlich und Künstlich unterteilt werden kann, oder sogar ein ganz eigenes Konstrukt zu Messungen bei Vertrauenswürdigkeit existiert.

6.2.1 Limitationen des Fragebogens

Obwohl die Ergebnisse aufgrund der Stichprobengröße sehr gut auf eine Mehrheit schließen lassen, besitzt der Fragebogen selbst jedoch einige Limitationen, beziehungsweise gibt es viele Aspekte, die von zukünftigen Forschenden untersucht werden können.

An erster Stelle steht hier das zuvor genannte neue Konstrukt der Messungen. Hierzu könnte ein ganz eigener Fragebogen erstellt werden, der nur auf Messungen und Messgeräten jeglicher Art eingeht und hier speziell die messenden Personen abdeckt.

Ein weiterer Schritt, der direkt bei den Ergebnissen des hier vorgestellten Fragebogens gesetzt werden kann, ist, jene Daten zu entfernen, die nicht sinnvoll sind. Einige Schüler*innen haben den Fragebogen nicht ernsthaft ausgefüllt und oft nur gleiche Antworten gegeben (obwohl darauf hingewiesen wurde, dass es sich hierbei um einen freiwilligen Fragebogen handelt). Würde man sich also nun genauer mit den Antworten auseinandersetzen und eben diese Fragebögen ausschließen, könnten die Antworten doch wieder anders aussehen.

Das eben angesprochene Problem hat natürlich auch Auswirkungen auf die hier gezeigten Ergebnisse. Gerade bei so jungen Schüler*innen kann nicht sicher gegangen werden, dass die Antworten gewissenhaft gegeben wurden. Alle Antworten sind also entsprechend mit Vorsicht zu genießen.

Grundsätzlich bildet dieser Fragebogen aber eine gute Grundlage für die Erforschung der Vertrauenswürdigkeit als Brückenkonzept zur Einführung von kleinen Messunsicherheiten in der SEKI.

7 Literatur

- Altrichter, H., Posch, P., & Spann, H. (2018). Lehrerinnen und Lehrer erforschen ihren Unterricht: Unterrichtsentwicklung und Unterrichtsevaluation durch Aktionsforschung. UTB.
- Bogner, A. (2002). Das Experteninterview: Theorie, Methode, Anwendung (A. Bogner, B. Littig, & W. Menz, Hrsg.; 2002. Aufl.). Vs Verlag Fur Sozialwissenschaften.
- Bärenthaler-Pachner, R. (2022). Entwicklung und Evaluation einer Lernumgebung zum Thema Messunsicherheit in der Sekundarstufe II (Masterarbeit, Universität Wien). Universität Wien.
- Cohen, J. (1960). 'A Coefficient of Agreement for Nominal Scales'. Educational and Psychological Measurement. Vol. 20, No. 1, 37-46.
- DATAtab Team (2023). DATAtab: Online Statistics Calculator. DATAtab e.U. Graz, Austria. URL <https://datatab.de/tutorial/fleiss-kappa> (zuletzt besucht am 18.05.2024)
- Drosg, M. (2006): Der Umgang mit Unsicherheiten. Ein Leitfaden zur Fehleranalyse. Wien: Facultas Universitätsverlag.
- Endreß, M. (2015). Der aktuelle theoretische Diskurs über Vertrauen in der Soziologie. Bielefeld: transcript Verlag
- Gehrke, M. (2022). Hauptkomponentenanalyse. In Angewandte Empirische Methoden in Finance and Accounting (pp. 334–351). Walter de Gruyter GmbH. <https://doi.org/10.1515/9783110767261-010>
- George, D., & Mallery, P. (2002). SPSS for windows step by step: A simple guide and reference, 11.0 update (4. Aufl.). Pearson.
- von Glasersfeld, E. (1997). Radikaler Konstruktivismus. Ideen, Ergebnisse, Probleme. Frankfurt am Main: Suhrkamp
- Groves, R. M. (2005). Survey errors and survey costs (1. Aufl.). John Wiley & Sons.
- Helfferrich, C. (2014). Leitfaden- und Experteninterviews. In N. Baur, J. Blasius (Hrsg.), Handbuch Methoden der empirischen Sozialforschung, DOI 10.1007/978-3-531-18939-0_39, Springer
- Hopf, M; Schecker, H.; Höttecke, D. und Wiesner, H. (2022) Physikdidaktik kompakt. Freising: Aulis-Verl
- Jordan, P. (2019). Faktorenanalyse (1. Auflage). Edition Rainer Hampp.

- Jung, W. (1986). Alltagsvorstellungen und das Lernen von Physik und Chemie. In R. Müller, R. Wodzinski & M. Hopf (Hrsg.), *Schülervorstellungen in der Physik* (S. 15 – 19). Aulis, 2007
- Kok, K., Priemer, B., Musold, W., & Masnick, A. (2019). Students' conclusions from measurement data: The more decimal places, the better? *Physical Review Physics Education Research*, 15(1). <https://doi.org/10.1103/physrevphyseducres.15.010103>
- Korner, Marianne.(2015). *Cross-Age Peer Tutoring in Physik: Evaluation einer Unterrichtsmethode*. Wien: Logos Verlag.
- Krüger, D., Parchmann, I., & Schecker, H. (Hrsg.). (2014). *Methoden in Der Naturwissenschaftsdidaktischen Forschung*. Springer.
- Landis, & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 159–174. <https://doi.org/10.2307/2529310>
- Loidl, H. (2021). *Entwicklung und Evaluation von Unterrichtseinheiten zum Thema Messunsicherheiten* (Masterarbeit, Universität Wien). Universität Wien.
- Luhmann, N. (2014). *Vertrauen - Ein Mechanismus der Reduktion sozialer Komplexität* (5.Aufl.). UTB.
- Nagel, C. (2021). Sicher ist sicher! Fachliche Klärung für die didaktische Rekonstruktion von Messunsicherheiten im Unterricht. *Plus Lucis* 1(4), 7-11.
- O'Connor, C., & Joffe, H. (2020). Intercoder Reliability in Qualitative Research: Debates and Practical Guidelines. *International Journal of Qualitative Methods*, 19,160940691989922. <https://doi.org/10.1177/1609406919899220>
- Osterloh, M. & Weibel, A. (2006). *Investition Vertrauen. Prozesse der Vertrauensentwicklung in Organisationen*. Gabler: Wiesbaden.
- Schendera, C. F. G. (2010). *Clusteranalyse mit SPSS: mit Faktorenanalyse*. Oldenbourg Verlag. <http://dx.doi.org/10.1524/9783486710526>
- Steiner, E., & Benesch, M. (2021). *Der Fragebogen: Von der Forschungsidee zur SPSS-Auswertung*. UTB.
- Vogl, S. (2022). Gruppendiskussion. In Baur, N., & Blasius, J. (Hrsg.). *Handbuch Methoden der empirischen Sozialforschung* (3. Aufl., S.913-919). Springer VS.
- Weichbold, M. (2014). Pretest. In N. Baur, J. Blasius (Hrsg.), *Handbuch Methoden der empirischen Sozialforschung*, DOI 10.1007/978-3-531-18939-0_39, Springer
- Wilson, M. (2004). *Constructing measures: An item response modeling approach*. Routledge Member of the Taylor and Francis Group.

A Erster Fragebogenentwurf

Informationsweitergabe

- Wenn eine Quelle **v.w. ist**, dann glaube ich jede Information, die von ihr kommt. (a)
- Halte ich eine Quelle **nicht für v.w.**, dann glaube ich keine einzige Information, die von ihr kommt. (a)
- Wenn eine Quelle **v.w. ist**, dann glaube ich fast jede Information, die von ihr kommt. (b)
- Wenn eine Quelle **nicht v.w. ist**, dann glaube ich fast keine einzige Information, die von ihr kommt. (b)
- Wenn eine Quelle **v.w ist**, dann wirkt sich das manchmal auf die Glaubwürdigkeit ihrer Informationen aus. (c)
- Wenn eine Quelle **nicht v.w ist**, dann wirkt sich das manchmal auf die Glaubwürdigkeit ihrer Informationen aus. (c)
- Wenn eine Quelle **v.w. ist**, dann wirkt sich das selten auf die Glaubwürdigkeit ihrer Informationen aus. (d)
- Wenn eine Quelle **nicht v.w. ist**, dann wirkt sich das selten auf die Glaubwürdigkeit ihrer Informationen aus. (d)
- Wenn eine Quelle **v.w. ist**, hat das keine Auswirkung auf die Glaubwürdigkeit ihrer Informationen. (e)
- Wenn eine Quelle **nicht v.w. ist**, hat das keine Auswirkung auf die Glaubwürdigkeit ihrer Informationen. (e)

Menschlich-künstlich

- Ob eine Quelle v.w. ist, hängt für mich immer vom Menschen dahinter ab. (a)
- Ob eine Quelle v.w. ist, hat nicht zwingend etwas mit dem Menschen dahinter zu tun. (d)
- Ob eine Quelle v.w. ist, hat nichts mit dem Menschen dahinter zu tun. (e)
- Ob eine Quelle v.w. ist, hängt ausschließlich von der Quelle selbst ab. (e)
- Ob eine Quelle v.w. ist, hängt hauptsächlich von der Quelle selbst ab. (d)
- Ob eine Quelle v.w. ist, hat nicht zwingend etwas mit der Quelle selbst zu tun, sondern mit dem Menschen dahinter. (b)
- Ob eine Quelle v.w. ist, hat nichts mit der Quelle selbst zu tun, sondern mit dem Menschen dahinter. (a)
- Ob eine Quelle v.w. ist, hängt gleichermaßen von dem Menschen dahinter und der Quelle selbst ab. (c)
- Wenn der Mensch hinter einer Quelle v.w. ist, dann ist es die Quelle automatisch auch. (a)

- Wenn der Mensch hinter einer Quelle v.w. ist, dann ist es die Quelle nicht automatisch auch. (d-e)
- Wenn eine Quelle v.w. ist, hat der Mensch dahinter keinen Einfluss darauf. (e)

Subjektiv - Objektiv

- Wenn eine andere, mir nahestehende, Person etwas für v.w. hält, dann gilt dies automatisch für mich.
- Wenn eine andere, mir fremde, Person etwas für v.w. hält, dann gilt dies automatisch für mich.
- Wenn eine andere, mir nahestehende, Person etwas für v.w. hält, dann gilt dies nicht automatisch für mich.
- Wenn eine andere, mir fremde, Person etwas für v.w. hält, dann gilt dies nicht automatisch für mich.
- Nur ich alleine entscheide für mich, was v.w. ist. (a)
- Wenn die Allgemeinheit etwas v.w. findet, so gilt dies automatisch für mich.
- Wenn die Allgemeinheit etwas v.w. findet, so gilt dies nicht automatisch für mich.
- Spricht die Allgemeinheit einer Quelle V.w. zu, so muss ich mir zuerst selbst ein Bild darüber machen.
- Spricht eine, mir nahestehende, Person einer Quelle V.w. zu, so muss ich mir zuerst selbst ein Bild darüber machen.
- Spricht eine, mir fremde, Person einer Quelle V.w. zu, so muss ich mir zuerst selbst ein Bild darüber machen.

Handelnde - Handlungen

- Wenn eine Person für mich v.w. ist, dann liegt das alleine an ihren Handlungen. (e)
- Wenn eine Person für mich v.w. ist, dann liegt das alleine an ihrer Art. (a)
- Ob eine Person v.w. ist, kann man alleine an ihren Handlungen messen. (e)
- Ob eine Person v.w. ist, kann man alleine an ihrem Charakter messen. (a)
- Auch wenn der Charakter einer Person v.w. ist, muss sie es zusätzlich durch Handlungen beweisen, damit dies für mich zählt. (c)
- Auch wenn die Handlungen einer Person v.w. scheinen, muss sie es zusätzlich durch ihre Art beweisen, damit dies für mich zählt. (c)
- V.w. hängt gleichermaßen von den Handlungen und der Art einer Person ab. (c)
- Auch wenn eine Person eine v.w. Art besitzt, so sind v.w. Handlungen ausschlaggebend für mich, um diese Person als v.w. zu bezeichnen. (d)

- Auch wenn eine Person v.w. handelt, so ist eine v.w. Art ausschlaggebend für mich, um diese Person als v.w. zu bezeichnen. (b)
- Auch wenn eine Person eine nicht v.w. Art besitzt, so kann sie durch v.w. Handlungen die Bezeichnung V.w. erlangen. (d-e)
- Trotz einer v.w. Art, gilt eine Person durch nicht v.w. Handlungen für mich als nicht v.w.. (d-e)
- Auch wenn eine Person nicht v.w. handelt, so kann sie durch v.w. Verhalten die Bezeichnung V.w. erlangen. (a-b)
- Trotz v.w. Handlungen, gilt eine Person durch eine nicht v.w. Art für mich als nicht v.w.. (a-b)

Wiederherstellung

- Die Art der Quelle hat immer einen Einfluss darauf, ob zerstörte V.w. wiederhergestellt werden kann.(a)
- Die Art der Quelle hat meistens einen Einfluss darauf, ob zerstörte V.w. wiederhergestellt werden kann.(b)
- Die Art der Quelle hat manchmal einen Einfluss darauf, ob zerstörte V.w. wiederhergestellt werden kann.(c)
- Die Art der Quelle hat selten einen Einfluss darauf, ob zerstörte V.w. wiederhergestellt werden kann.(d)
- Die Art der Quelle hat nie einen Einfluss darauf, ob zerstörte V.w. wiederhergestellt werden kann.(e)

Schwere des Vertrauensbruches

- Die Schwere des Vertrauensbruches hat immer einen Einfluss darauf, ob die zerstörte V.w. wiederhergestellt werden kann.(a)
- Die Schwere des Vertrauensbruches hat meistens einen Einfluss darauf, ob die zerstörte V.w. wiederhergestellt werden kann.(b)
- Die Schwere des Vertrauensbruches hat manchmal einen Einfluss darauf, ob die zerstörte V.w. wiederhergestellt werden kann.(c)
- Die Schwere des Vertrauensbruches hat selten einen Einfluss darauf, ob die zerstörte V.w. wiederhergestellt werden kann.(d)
- Die Schwere des Vertrauensbruches hat nie einen Einfluss darauf, ob die zerstörte V.w. wiederhergestellt werden kann.(e)

(Un)verzichtbar

- Mir ist es wichtig, dass Quellen, mit denen ich zu tun habe, auf mich v.w. wirken.(a)

- Mir ist es wichtig, dass Quellen, mit denen ich zu tun habe, mir ihre V.w. beweisen.(a)
- Mir ist es egal, ob Quellen, mit denen ich zu tun habe, auf mich v.w. wirken. (e)
- Mir ist es egal, ob Quellen, mit denen ich zu tun habe, mir ihre V.w. beweisen. (e)
- Mir ist es nicht immer wichtig, dass Quellen, mit denen ich zu tun habe, mir ihre V.w. beweisen. (c)
- Mir ist es nicht immer wichtig, dass Quellen, mit denen ich zu tun habe, auf mich v.w. wirken. (c)
- Mir ist es fast immer wichtig, dass Quellen, mit denen ich zu tun habe, mir ihre V.w. beweisen. (b)
- Mir ist es fast immer wichtig, dass Quellen, mit denen ich zu tun habe, auf mich v.w. wirken. (b)
- Mir ist es fast nie wichtig, ob Quellen, mit denen ich zu tun habe, mir ihre V.w. beweisen. (d)
- Mir ist es fast nie wichtig, ob Quellen, mit denen ich zu tun habe, auf mich v.w. wirken. (d)

B Überarbeitung: Erster Fragebogenentwurf

	trifft zu	trifft eher zu	teils-teils	trifft eher nicht zu	trifft nicht zu
Eine Instagram-Seite, die von mehreren Personen betrieben wird, postet einen Beitrag. Ob dieser Beitrag auf mich vertrauenswürdig wirkt, hängt von den einzelnen Personen ab.	☐	☐	☐	☐	☐
Eine Firma besitzt eine Internetseite, auf welcher regelmäßig Beiträge geteilt werden. Die Vertrauenswürdigkeit der Seite hat nichts damit zu tun, ob die Personen hinter der Seite vertrauenswürdig sind.	☐	☐	☐	☐	☐
Eine für mich vertrauenswürdige Influencerin besitzt einen TikTok-Account. Da die Influencerin vertrauenswürdig ist, ist es der Account ebenfalls.	☐	☐	☐	☐	☐
Ob eine Instagram-Seite vertrauenswürdig ist, hängt nicht von der Vertrauenswürdigkeit von dem Creator dahinter ab.	☐	☐	☐	☐	☐
Wenn ich eine Instagram-Seite für vertrauenswürdig halte, dann trifft dies automatisch auf alle Informationen zu, die diese Seite teilt.	☐	☐	☐	☐	☐
Wirkt ein Influencer auf mich nicht vertrauenswürdig, dann sind automatisch alle Informationen, die dieser Influencer teilt, nicht vertrauenswürdig.	☐	☐	☐	☐	☐
Ich sehe ein Video auf einem TikTok-Account. Auch wenn der Account auf mich vertrauenswürdig wirkt, sagt das nichts über die Vertrauenswürdigkeit des Videos selbst aus.	☐	☐	☐	☐	☐
Ich sehe einen Post auf einer Instagram-Seite. Auch wenn die Instagram-Seite auf mich nicht vertrauenswürdig wirkt, hat dies keinen Auswirkung auf die Vertrauenswürdigkeit des Posts selbst.	☐	☐	☐	☐	☐
Eine gute Freundin zeigt mir einen TikTok-Account. Da sie diesen vertrauenswürdig findet, gilt dies auch automatisch für mich.	☐	☐	☐	☐	☐
Eine mir fremde Person empfiehlt mir eine vertrauenswürdige Instagram-Seite. Da die Person die Seite für vertrauenswürdig hält, gilt dies auch automatisch für mich.	☐	☐	☐	☐	☐
Wenn ich einen neuen TikTok-Account entdecke, muss ich mir zuerst immer selbst ein Bild davon machen und kann nur damit entscheiden, ob dieser vertrauenswürdig ist.	☐	☐	☐	☐	☐
Es gibt Instagram-Seiten, bei denen von vornherein klar ist, dass diese vertrauenswürdig sind, da dies allgemein gültig ist.	☐	☐	☐	☐	☐
Auch wenn eine Internetseite allgemein als vertrauenswürdig gilt, muss ich mir diese Seite trotzdem anschauen. Erst dann kann ich sagen, ob sie vertrauenswürdig ist, oder nicht.	☐	☐	☐	☐	☐
Ein guter Freund erzählt mir von einer neuen Instagram-Seite, die er für vertrauenswürdig hält. Auch wenn er so denkt, muss ich mir zuerst selbst ein Bild davon machen, um die Seite für vertrauenswürdig zu halten.	☐	☐	☐	☐	☐
Jemand auf der Straße empfiehlt mir einen neuen TikTok-Account, den die Person für vertrauenswürdig hält. Obwohl die Person den Account für vertrauenswürdig hält, muss ich mir selbst ein Bild dazu machen, damit ich diesen Account als vertrauenswürdig einstufen kann.	☐	☐	☐	☐	☐
Eine gute Bekannte erzählt mir von einem nicht vertrauenswürdigem TikTok-Account. Da sie diesen nicht für vertrauenswürdig hält, gilt dies auch automatisch für mich.	☐	☐	☐	☐	☐
Eine mir fremde Person warnt mich vor einer nicht vertrauenswürdigem Instagram-Seite. Da die Person die Seite für nicht vertrauenswürdig hält, gilt dies auch automatisch für mich.	☐	☐	☐	☐	☐
Eine bestimmte Internetseite gilt allgemein als nicht vertrauenswürdig. Dies gilt dann auch automatisch für mich.	☐	☐	☐	☐	☐
Ein bestimmter TikTok-Account wird allgemein als nicht vertrauenswürdig bezeichnet. Trotz dieser Tatsache muss ich mir zuerst selbst ein Bild davon, um den Account beurteilen zu können.	☐	☐	☐	☐	☐

Ein guter Bekannter erzählt mir von einer nicht vertrauenswürdigen Instagram-Seite. Obwohl er die Seite für nicht vertrauenswürdig hält, muss ich mir selbst ein Bild von der Seite machen, um dies beurteilen zu können.	☐	☐	☐	☐	☐
Eine mir fremde Person weißt mich auf einen nicht vertrauenswürdigen TikTok-Account hin. Auch wenn die Person den Account für nicht vertrauenswürdig hält, muss ich mir zuerst selbst ein Bild davon machen, damit ich dies beurteilen kann.	☐	☐	☐	☐	☐
Eine Person in meinem Umfeld muss vertrauenswürdige Dinge tun, um auf mich vertrauenswürdig zu wirken.	☐	☐	☐	☐	☐
Damit eine Person auf mich vertrauenswürdig wirkt, muss sie eine vertrauenswürdige Art besitzen.	☐	☐	☐	☐	☐
Eine Person muss vertrauenswürdige Sachen sagen, um auf mich vertrauenswürdig zu wirken, auch wenn sie eine vertrauenswürdige Art besitzt.	☐	☐	☐	☐	☐
Eine Person in meinem Umfeld muss eine vertrauenswürdige Art besitzen, um auf mich vertrauenswürdig zu wirken, auch wenn sie vertrauenswürdige Dinge macht.	☐	☐	☐	☐	☐
Mir ist es wichtig, dass ein TikTok-Account, den ich regelmäßig schaue, auf mich vertrauenswürdig wirkt.	☐	☐	☐	☐	☐
Wenn ich einer Instagram-Seite folge und ihre Beiträge regelmäßig ansehe, dann ist es für mich nicht wichtig, ob die Seite auf mich vertrauenswürdig wirkt.	☐	☐	☐	☐	☐
Vertrauenswürdigkeit ist für mich eine positive Eigenschaft.	☐	☐	☐	☐	☐
Vertrauenswürdigkeit ist für mich eine negative Eigenschaft.	☐	☐	☐	☐	☐
Vertrauenswürdigkeit und Verlässlichkeit sind für mich das Gleiche.	☐	☐	☐	☐	☐
Vertrauenswürdigkeit und Aufrichtigkeit sind für mich das Gleiche.	☐	☐	☐	☐	☐
Vertrauenswürdigkeit und geringe Messunsicherheit sind für mich das Gleiche.	☐	☐	☐	☐	☐
Vertrauenswürdigkeit und Unglaubwürdigkeit sind für mich etwas anderes.	☐	☐	☐	☐	☐
Vertrauenswürdigkeit und Unehrllichkeit sind für mich etwas anderes.	☐	☐	☐	☐	☐
Vertrauenswürdig und unseriös sind für mich etwas anderes.	☐	☐	☐	☐	☐
Vertrauenswürdig und unzuverlässig sind für mich das Gleiche.	☐	☐	☐	☐	☐
Vertrauenswürdigkeit und Fröhlichkeit sind für mich das Gleiche.	☐	☐	☐	☐	☐
Vertrauenswürdigkeit und Faulheit sind für mich etwas anderes.	☐	☐	☐	☐	☐
	☐	☐	☐	☐	☐
Wenn ich auf einem vertrauenswürdigen Instagram-Account von Ergebnissen einer Studie lese, so sind diese ebenfalls vertrauenswürdig.	☐	☐	☐	☐	☐
Im Physikunterricht wird bei einem Versuch die Zeit mit dem Sekundenzeiger einer Uhr gestoppt. Dieses Messergebnis wäre vertrauenswürdiger, wenn man die Zeit mit dem Smartphone gestoppt hätte.	☐	☐	☐	☐	☐
Da ich die Antwort auf eine Aufgabe nicht weiß, gebe ich diese Aufgabe in einen Chatbot ein. Die Antwort von diesem Chatbot schätze ich als vertrauenswürdig ein.	☐	☐	☐	☐	☐
In der Schule findet die Wahl zum*zur Schüler*innensprecher*in statt. Die Wahlzettel werden anschließend von drei vertrauenswürdigen Personen einmal gezählt. Das Ergebnis dieser Wahl wirkt auf mich vertrauenswürdig.	☐	☐	☐	☐	☐
Wörter lassen sich mit Google Übersetzer von Deutsch auf Englisch übersetzen. Ich finde, dass diese Übersetzungen vertrauenswürdig sind.	☐	☐	☐	☐	☐
Wenn ich etwas auf Wikipedia lese, so schätze ich diese Information als vertrauenswürdig ein.	☐	☐	☐	☐	☐

Ich finde eine Anleitung zu einer Schulaufgabe, die von einem Chatbot geschrieben wurde. Ich übernehme diese Anleitung, ohne mir diese genauer durchzulesen, da sie auf mich vertrauenswürdig wirkt.	☐	☐	☐	☐	☐
In der Schule wird eine Umfrage zum Thema "Essgewohnheiten" durchgeführt. Die Daten werden anschließend von einem Programm ausgewertet. Die Ergebnisse dieser Umfrage wirken vertrauenswürdig auf mich.	☐	☐	☐	☐	☐
Im Physikunterricht führen wir einen Versuch durch, bei dem ein Ball fallen gelassen und die Zeit dazu gestoppt wird. Die Zeit wird hier von einem Gerät gestoppt, welches mit Lichtschranken sieht, wann der Ball fallen gelassen und wann er am Boden aufkommt. Die Messergebnisse halte ich für vertrauenswürdig.	☐	☐	☐	☐	☐

C Zweiter Fragebogenentwurf

	sehr ähnlich	sehr verschieden	haben nichts miteinander zu tun
Unsicherheit und Vertrauenswürdigkeit sind für mich...	☐	☐	☐
Richtigkeit und Vertrauenswürdigkeit sind für mich...	☐	☐	☐
Glaubwürdigkeit und Vertrauenswürdigkeit sind für mich...	☐	☐	☐
Faulheit und Vertrauenswürdigkeit sind für mich...	☐	☐	☐
Verlässlichkeit und Vertrauenswürdigkeit sind für mich...	☐	☐	☐
Unehrlichkeit und Vertrauenswürdigkeit sind für mich...	☐	☐	☐
Unglaubwürdigkeit und Vertrauenswürdigkeit sind für mich...	☐	☐	☐
Langeweile und Vertrauenswürdigkeit sind für mich...	☐	☐	☐
Zwielichtigkeit und Vertrauenswürdigkeit sind für mich...	☐	☐	☐
Aufrichtigkeit und Vertrauenswürdigkeit sind für mich...	☐	☐	☐
Geringe Messunsicherheit und Vertrauenswürdigkeit sind für mich...	☐	☐	☐

	trifft zu	trifft eher zu	teils-teils	trifft eher nicht zu	trifft nicht zu
Ich halte Informationen in meinem Schulbuch für vertrauenswürdig.	☐	☐	☐	☐	☐
Ich halte Informationen von meiner Lehrperson für vertrauenswürdig.	☐	☐	☐	☐	☐
Ich halte Informationen von meinen Eltern für vertrauenswürdig.	☐	☐	☐	☐	☐
Ich halte Informationen von meinem/meiner Nachhilfelehrer*in für vertrauenswürdig.	☐	☐	☐	☐	☐
Ich halte Informationen von Wikipedia für vertrauenswürdig.	☐	☐	☐	☐	☐
Ich halte Informationen von YouTube-Lernvideos für vertrauenswürdig.	☐	☐	☐	☐	☐
Ich halte Informationen von unbekannten Anrufern für vertrauenswürdig.	☐	☐	☐	☐	☐
Ich halte Informationen aus dem Internet für vertrauenswürdig.	☐	☐	☐	☐	☐
Ich halte Informationen auf Sozialen Medien (Instagram, TikTok,...) für vertrauenswürdig.	☐	☐	☐	☐	☐
Ich halte die Ergebnisse des Taschenrechners für vertrauenswürdig.	☐	☐	☐	☐	☐
Ich halte die Ergebnisse im Lösungsheft für vertrauenswürdig.	☐	☐	☐	☐	☐
Ein*e Freund*in und ich stoppen die Zeit bei einem Wettlauf. Ich vertraue meinem Messergebnis mehr als seinem/ihrer.	☐	☐	☐	☐	☐
Ein*e Wissenschaftler*in und ich stoppen die Zeit bei einem Wettlauf. Ich vertraue meinem Messergebnis mehr als seinem/ihrer.	☐	☐	☐	☐	☐
Eine Lehrperson und ich stoppen die Zeit bei einem Wettlauf. Ich vertraue meinem Messergebnis mehr als seinem/ihrer.	☐	☐	☐	☐	☐
Ich wiege beim Backen das Mehl ab. Die Waage zeigt minus 10 Gramm an. Ich halte dieses Messergebnis für vertrauenswürdig.	☐	☐	☐	☐	☐
Ich wiege eine Kilopackung (1000 Gramm) Mehl auf einer Waage ab. Die Waage zeigt 999g an. Ich halte dieses Messergebnis für vertrauenswürdig.	☐	☐	☐	☐	☐
Mit einem Lineal kann ich vertrauenswürdig eine Länge von einem Zentimeter messen.	☐	☐	☐	☐	☐
Draußen schneit es. Das Außenthermometer zeigt 40°C an. Ich halte dieses Messergebnis für vertrauenswürdig.	☐	☐	☐	☐	☐
Eine Schulärztin misst am Montag meine Körpergröße, das Messergebnis ist 159,0 cm. Am Dienstag misst sie nochmal. Nun ist das Ergebnis 158,5 cm. Ich halte dieses Messergebnis für vertrauenswürdig.	☐	☐	☐	☐	☐

Ein YouTube-Kanal dem ich vertraue, wechselt den Content-Creator (Autor). Ich vertraue dem Kanal weiterhin.	☐	☐	☐	☐	☐
Mein Großvater kann mir nicht die 9 Bundesländer Österreichs nennen. Ich vertraue ihm trotzdem.	☐	☐	☐	☐	☐
Die Antworten die mir ChatGPT (ChatBot) auf eine Frage gibt, halte ich (grundsätzlich) für vertrauenswürdig.	☐	☐	☐	☐	☐
Ich halte fremde Personen, die mir am Bahnsteig irgendwelche Waren anbieten wollen, für vertrauenswürdig.	☐	☐	☐	☐	☐

D Informationen an die Erziehungsberechtigten

Sehr geehrte Eltern,

wir, eine Gruppe von Studierenden der Universität Wien, schreiben Ihnen heute, um Sie über unsere bevorstehende Forschungsarbeit zum Thema Vertrauenswürdigkeit bei jungen Menschen zu informieren. Im Rahmen unserer Studie werden wir Interviews und Umfragen in den Unterstufen durchführen, um ein besseres Verständnis für die Wahrnehmung von Vertrauen und Vertrauenswürdigkeit bei Schüler*innen zu erhalten.

Wir möchten Ihnen versichern, dass die Privatsphäre und die Vertraulichkeit Ihrer Kinder während der gesamten Studie gewahrt bleiben werden. Bevor wir mit den Interviews und Umfragen beginnen, werden wir eine Einverständniserklärung von Ihnen einholen, um sicherzustellen, dass Sie über unsere Studie informiert sind und zustimmen, dass Ihre Kinder daran teilnehmen dürfen.

Wir werden uns an alle geltenden Richtlinien und Bestimmungen halten und sicherstellen, dass alle Informationen, die wir sammeln, anonym und vertraulich bleiben. Wir möchten betonen, dass die Teilnahme an unserer Studie freiwillig ist und Ihre Kinder jederzeit das Recht haben, ihre Teilnahme abzuberechnen.

Wir sind der Meinung, dass diese Forschungsarbeit von großem Nutzen sein wird, um ein besseres Verständnis für die Wahrnehmung von Vertrauen und Vertrauenswürdigkeit bei jungen Menschen zu gewinnen. Wir möchten uns im Voraus bei Ihnen bedanken, dass Sie uns erlauben, unsere Forschungsarbeit in Ihrer Schule durchzuführen.

Mit freundlichen Grüßen,

Eva Abl, Nicolas Hoang, Raphaela Pelz, Tom Schnögl, Betül Akalin

E Einverständniserklärung

Einverständniserklärung zur Teilnahme an Befragungen und Interviews

Ich erkläre hiermit mein Einverständnis zur Teilnahme meines Kindes an Befragungen und Interviews im Rahmen einer Forschungsarbeit zum Thema Vertrauenswürdigkeit, die von einer Gruppe von Studierenden der Universität Wien durchgeführt wird.

Ich verstehe, dass die Befragungen und Interviews freiwillig sind, und dass mein Kind das Recht hat, jederzeit während des Prozesses abubrechen.

Ich bin damit einverstanden, dass mein Kind an den Befragungen und Interviews teilnimmt und dass die Informationen, die mein Kind bereitstellt, anonym und vertraulich behandelt werden. Ich stimme auch zu, dass die Studierendengruppe die Ergebnisse ihrer Forschungsarbeit veröffentlichen darf, jedoch ohne dass dabei mein Kind oder seine Identität offengelegt wird.

Name des Elternteils: _____

Name des Kindes: _____

Unterschrift des Elternteils: _____

Datum: _____