Der Umgang mit Daten aus erster und zweiter Hand im Physikunterricht

Stephan Pfeiler, Burkhard Priemer

Humboldt-Universität zu Berlin, Institut für Physik, Newtonstraße 15, 12489 Berlin
pfeilers@physik.hu-berlin.de, priemer@physik.hu-berlin.de

Kurzfassung

Der Umgang mit und die Evaluation von Daten ist Teil wissenschaftlichen Arbeitens. Um Fertigkeiten im Umgang mit Daten im Unterricht zu vermitteln, werden in der Regel Daten aus Experimenten verwendet. Diese können von den Schülerinnen und Schülern selbst (Daten aus erster Hand) oder von anderen Personen (Daten aus zweiter Hand) erhoben werden. Die Akzeptanz von Daten kann als Ergebnis einer Einschätzung der Glaubwürdigkeit dieser Daten oder des Überbringers der Daten betrachtet werden. Das Konstrukt der Glaubwürdigkeit kann somit ein Werkzeug sein, um die unterschiedliche Wirkung verschiedener Datentypen zu verstehen. Diese Arbeit stellt eine Interviewstudie vor, welche die Nutzung von Kriterien für die Bewertung von Glaubwürdigkeit durch Schülerinnen und Schüler untersuchte, nachdem diese entweder mit Daten aus erster Hand oder Daten aus zweiter Hand konfrontiert wurden. Die Studie konnte keine nennenswerten, signifikanten Unterschiede zwischen den Versuchsgruppen ausmachen. Das gilt sowohl für die Nutzung der Kriterien durch die Probanden als auch für ein Ranking von vorgegebenen Aspekten der Glaubwürdigkeitsbewertung.

1. Theorie

Daten sind ein zentraler Bestandteil von naturwissenschaftlichem Unterricht. Sie sind in verschiedensten Modellen des wissenschaftlichen Denkens zentrales Element (Klahr, 2002; Koslowski, 1996; Kuhn, 2002). Daten oder Evidenz werden genutzt, um Entscheidungen für oder gegen bestimmte Vorstellungen zu untermauern. Sie dienen der Koordination von Theorien über die Welt mit den Beobachtungen in dieser Welt. In einfachen Modellen dienen Daten zur Entscheidung zwischen konkurrierenden Hypothesen (Chinn & Brewer, 1993). Diese Hypothesen sind Vorhersagen darüber wie die Natur auf Veränderungen reagiert, basierend auf Theorien über die Funktionsweisen der Natur. Eine Grundlage der Entscheidung für oder gegen eine Hypothese ist die Interpretation der Daten bezüglich der Hypothese. Außerdem müssen Daten evaluiert werden. Nicht jede Interpretation von Daten ist berechtigt. Es kann verschiedenste Gründe für die Ablehnung von Daten geben. So wird die Evaluation von Daten ebenfalls als wichtiger Bestandteil des Umgangs mit Daten verstanden. Im SDDS-Modell entspricht die Evaluation von Evidenz einer von drei zentralen Tätigkeiten im wissenschaftlichen Prozess (Klahr, 2002). Leider spielt die Evaluation von Daten im Physikunterricht eine untergeordnete Rolle. Auch der Bezug zu Hypothesenentscheidungen fehlt in der Regel (Tesch, 2005). Schülerinnen und Schüler lernen nur in seltenen Fällen den Umgang mit Messunsicherheiten (Heinicke, 2012).

Im Physikunterricht spielen immer wieder Daten verschiedenen Typs eine Rolle. Zwei Grundformen des Experiments im Unterricht sind das Demonstrations- und das Schülerexperiment. Des Weiteren können Daten aus anderen Quellen herangezogen werden. Dies führt zu einer Unterscheidung in Daten aus erster Hand und Daten aus zweiter Hand (Hug & McNeill, 2008). Daten aus erster Hand sind im eigenen Experiment gewonnene Daten. Daten aus zweiter Hand stammen aus anderen Quellen.

Es ist von didaktischer Bedeutung zu wissen, mit welchen alltäglichen Einstellungen Schülerinnen und Schüler an die Evaluation von Daten aus unterschiedlichen Quellen herantreten. Welche Kriterien nutzen die Schülerinnen und Schüler, um zu entscheiden, ob Daten in einem physikalischen Kontext für oder gegen eine Hypothese sprechen? Zur Beschreibung des Prozesses der Entscheidung für oder gegen einen Hypothesenwechsel nutzen wir in dieser Arbeit das Konstrukt der Glaubwürdigkeit von Daten. Glaubwürdigkeit taucht in verschiedenen Arbeiten zur Evaluation von Daten auf (Chinn & Brewer, 2001; Driver, Newton, & Osborne, 2000). Oft bleibt der Begriff lediglich auf Personen beschränkt. Eine umfassendere, aber weniger spezifische Definition von Glaubwürdigkeit versteht diese als Eigenschaft, die sowohl Personen, Organisationen sowie ihren kommunikativen Produkten zugewiesen werden kann. Dies umfasst alle an der Evaluation von Daten beteiligten Akteure: den Autor der Daten sowie die Daten selbst (als kommunikatives Produkt). Welche Kriterien für oder gegen die Glaubwürdigkeit sprechen ist in hohem Maße vom Kontext abhängig, in dem diese Glaubwürdigkeit bewertet werden soll (Rieh & Danielson, 2007). Glaubwürdigkeit dient in

diesem Verständnis als Konstrukt zur Beschreibung der Evaluation von Daten aus unterschiedlichen Quellen durch Schülerinnen und Schüler. Dieses Konstrukt gilt es auf den Physikunterricht zu beziehen und zu spezifizieren.

2. Fragestellung

Es ergeben sich die folgenden Forschungsfragen:

a) Welchen Einfluss haben Daten aus erster und Daten aus zweiter Hand auf die Entscheidung für oder gegen eine eingangs aufgestellte Hypothese?

b) Welchen Einfluss haben Daten aus erster und Daten aus zweiter Hand auf die Verwendung von Kriterien für die Bewertung der Glaubwürdigkeit dieser Daten?

3. Vorarbeiten

Im Rahmen einer Vorstudie wurden 19 Schülerinnen und Schüler zur Glaubwürdigkeit von verschiedenen Datensätzen zum Zusammenhang zwischen schwingender Masse und Periodendauer eines Fadenpendels befragt. Die Aussagen der Schülerinnen und Schüler wurden hinsichtlich der von ihnen genannten Kriterien analysiert und kategorisiert. Auf diese Weise ist ein Codesystem bestehend aus 4 Codes und jeweils 5-6 Subcodes entstanden. Die Codes lauten „Eigenschaften des Experiments“, „Eigenschaften von Autoren“, „Eigenschaften der Daten“ und „Prüfen und Abgleichen“. Die drei erstgenannten Codes konnten bereits vor der Untersuchung deduktiv abgeleitet werden. Sie entsprechen den von einer Glaubwürdigkeitsbewertung betroffenen Aspekten: Daten und die damit verbundene Methodik der Datenerhebung sowie der Quelle der Daten. Damit ist in diesem Fall der Autor gemeint (eine Systematik von Kriterien für die Glaubwürdigkeit wurde auf Basis der folgenden Quellen erarbeitet: Bentele, Brosius, & Jarren, 2012; Driver, Newton, & Osborne, 2000; Nicolaidou, Kyza, Terzian, Hadjichambis, & Kafouris, 2011; Rieh & Danielson, 2007; Wilson, 1983) Der vierte Code „Prüfen und Abgleichen“ umfasst Aussagen, welche sich nicht mit einer spezifischen Charakteristik des Datensatzes auseinandersetzen, sondern Daten Glaubwürdigkeit zusprechen, wenn sie überprüft oder mit Referenzen verglichen wurden.

4. Methode

4.1. Datenerhebung

In der Studie haben insgesamt 21 Schülerinnen und 21 Schüler der neunten Klassenstufe an drei Berliner Gymnasien im Alter von 14 bis 16 Jahren (MW: 14,2 Jahre) Messdaten zu einem Experiment mit dem Fadenpendel aufgenommen. Sie untersuchten den Zusammenhang zwischen der Schwingungsdauer des Pendels und der schwingenden Masse. Dieser Versuch wurde gewählt, da er gezielt Alltagsvorstellungen der Schülerinnen und Schüler aufgreift, wonach schwerere Massen schneller schwingen, also eine kleinere Schwingungsdauer aufweisen. Die entstandenen Daten sind aus Sicht vieler Schülerinnen und Schüler deshalb nicht erwartungskonform. Der Versuch ist zudem vergleichsweise einfach durchzuführen und generiert Datensätze in angemessener Qualität (Kanari & Millar, 2004). Diese dienen dazu die eigenen Vorstellungen in Frage zu stellen (Chinn & Brewer, 1993) und werden hier als Gesprächsanlass genutzt. Vor dem Experiment entschied sich jeder Proband für eine von drei möglichen Hypothesen bzgl. der Periodendauer des Fadenpendels: Bei steigender Masse wird die Schwingungsdauer (a) kleiner, (b) bleibt gleich oder (c) wird größer. Danach nahmen die Teilnehmer zu drei verschiedenen Massen jeweils 10 Messwerte auf, sodass jedem Probanden ein Datensatz mit insgesamt 30 Einzelmessungen zur Verfügung stand. Zusätzlich zu den eigenhändig aufgenommenen Daten standen zwei vorbereitete Datensätze im gleichen Umfang zur Verfügung. Sie bestanden aus den gleichen Einzelmessungen und wurden randomisiert angeordnet, sodass dies für die Probanden nicht ersichtlich war. Diese beiden Datensätzen wurden einer Teilstichprobe der Probanden mit dem Hinweis vorgelegt, dass die Autoren der Daten „Schüler“ bzw. „Lehrer“ sind. „Schüler“ bedeutete, dass die Daten von einem fiktiven Schüler der neunten Klasse eines anderen Berliner Gymnasiums stammten. „Lehrer“ bedeutete, dass die Daten von einem fiktiven Physiklehrer eines anderen Berliner Gymnasiums stammten. So ergaben sich drei Versuchsgruppen à 14 Probanden. Die erste Versuchsgruppe behielt ihre eigenen Daten. In der zweiten und dritten Versuchsgruppe wurden die eigenen Daten durch einen der beiden Datensätze „Schüler“ oder „Lehrer“ ersetzt.

Im Anschluss an die Messungen wurden alle Probanden in einem halbstrukturierten Interview befragt, ob die vorliegenden Daten die eingangs aufgestellte Hypothese bestätigen oder widerlegen. Es wurde gezielt nach der Glaubwürdigkeit des Datensatzes gefragt, wobei Wert darauf gelegt wurde, dass die Nennung von Kriterien für die Glaubwürdigkeit durch die Interviewten erfolgte. Die Rolle des Interviewers beschränkte sich auf die Motivierung der Interviewten, ihre Aussagen zu ergänzen oder zu spezifizieren. Zum Abschluss des Interviews wurden die Interviewten gebeten ein Rating der vier Codes aus der Vorstudie vorzunehmen. Das Rating bestand darin, die Codes in eine Rangfolge zu bringen, wobei die Bedeutung der Codes als „Für die Bewertung der Glaubwürdigkeit am wichtigsten“ bis hin zu „Für die Bewertung der Glaubwürdigkeit am unwichtigsten“ beschrieben wurde.

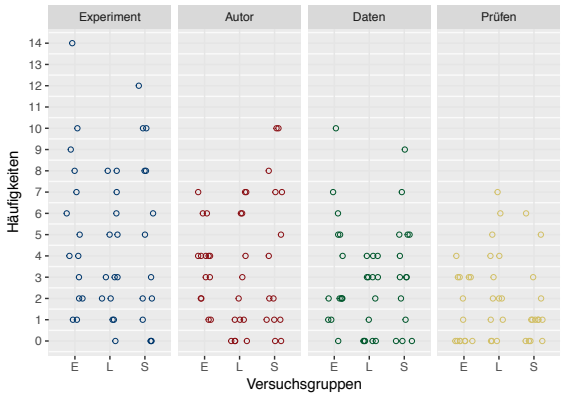

Abb. 1: Häufigkeitsverteilungen von Aussagen in den Codes. Jeder Punkt symbolisiert ein Interview. Die vertikale Position gibt die Anzahl der Aussagen an, welche in diesem Interview mit dem jeweiligen Code codiert wurden. E bezeichnet Probanden mit eigenen Daten, L bezeichnet Probanden mit Lehrerdaten, S bezeichnet Probanden mit Schülerdaten.

Abb. 2: Verteilung der Häufigkeiten ausgewählter Subcodes (Durchführung (13), Menschentoleranz (16), Fehlbarkeit (26), Streuung (36), Abgleich mit eigenen Daten (42))

4.2. Auswertung

Die Interviews wurden transkribiert. Auf Basis eines Regelkatalogs wurden anschließend Aussagen für die Codierung aus der Vorstudie ausgewählt und in Codiereinheiten vorstrukturiert. Zwei unabhängige Rater codierten in zwei Durchläufen jeweils 9 Interviews und konnten im Gespräch eventuelle Unstimmigkeiten ausräumen und das Codiermanual anpassen. Die Beurteilerübereinstimmung wurde sowohl auf Ebene der Codes als auch auf Ebene der Subcodes bestimmt. Für die Subcodes griffen wir auf ein gewichtetes Kappa κ_w (Cohen, 1968) zurück. Dies erlaubt fehlende Übereinstimmungen zwischen Subcodes verschiedener Codes unterschiedlich stark in den Wert der Beurteilerübereinstimmung einzubeziehen. Für die Beurteilerübereinstimmung auf Ebene der Codes wurde die Gewichtungsmatrix des κ_w so angepasst, dass ein ungewichtetes κ für die 4 Codes berechnet werden konnte, ohne das Interviewmaterial erneut zu codieren. Nach Gwet (2012) wurden die κ-Werte angepasst, um eine realistischere Abschätzung der Beurteilerübereinstimmung unter Berücksichtigung von Randbedingungen (Anzahl der beurteilten Aussagen, Anzahl der Rater, Anzahl der Kategorien) zu gewährleisten. Die Beurteilerübereinstimmung auf Code-Ebene beträgt κ_w = .66 . Dies entspricht einer substantiellen Übereinstimmung (Landis & Koch, 1977). Auf Subcode-Ebene beträgt die Beurteilerübereinstimmung κ = .59 . Dies entspricht einer moderaten Übereinstimmung (ebenda). Die restlichen Interviews wurden anschließend von einem der beiden Rater codiert.

Inwiefern die Daten einen Einfluss auf den Hypothesenwechsel der Probanden hatten, wurde festgestellt, indem Aussagen identifiziert wurden, die einen Rückschluss darüber zuließen, ob der Proband seine eingangs aufgestellte Hypothese im Licht der neuen Daten beibehält oder nicht. Im Falle eines Hypothesenwechsels wurde festgehalten, für welche neue Hypothese sich die Probanden entscheiden. Dies war für 39 der 42 Interviews möglich. Ob signifikante Unterschiede zwischen den verschiedenen Wechselmöglichkeiten und Versuchsgruppen bestanden, wurde mit $k \times l$-χ^2-Test unter Zuhilfenahme von Fishers exaktem Test (auf Grund zu geringer Randsummen) berechnet (Bortz & Schuster, 2010; Field & Miles, 2012).

Alle Aussagen der Probanden, in denen die Glaubwürdigkeit angesprochen wurde, wurden mit dem Codesystem codiert. So konnten Häufigkeiten bezüglich der Verwendung einzelner Subcodes und der Codes ermittelt werden. Aufgrund der nur moderaten Beurteilerübereinstimmung auf Subcode-Ebene erfolgte eine Bestimmung von Signifikanzen bezüglich der Unterschiede der Häufigkeiten in Abhängigkeit von der Versuchsgruppe nur auf Ebene der Codes. Dafür wurde der Kruskal-Wallis-Test (Field & Miles, 2012) verwendet.

Auf Subcode-Ebene wird die Analyse von Codierungshäufigkeiten auf die fünf häufigsten Subcodes beschränkt. Die Analyse von Unterschieden zwischen den Versuchsgruppen wurde ebenfalls mit dem Kruskal-Wallis-Test durchgeführt.

Die Ergebnisse aus dem Ranking der Codes nach Wichtigkeit durch die Probanden wurden numerisch festgehalten. Für jeden Probanden erhielten die gerankten Codes einen von vier Rängen (1→am wichtigsten bis 4→am unwichtigsten) und konnten so analysiert werden. Unterschiede zwischen den Häufigkeiten konnten auch hier mit Hilfe des Kruskal-Wallis-Tests ermittelt werden.

5. Ergebnisse

5.1. Hypothesenwechsel

Diese Studie konnte frühere Ergebnisse zum Fadenpendel reproduzieren (Kanari & Millar, 2004; Ludwig & Priemer, 2013). Von den 39 Probanden, deren Hypothesenwechsel nachvollzogen werden konnte, entschieden sich 34 für die fachlich falsche

Hypothese. Nach dem Experiment und der Konfrontation mit einem der verschiedenen Datensätze wechselten 18 Probanden von einer fachlich falschen Hypothese zu der fachlich richtigen Hypothese. 16 Probanden behielten die fachlich falsche Hypothese bei oder wechselten zu der anderen fachlich falschen Hypothese. Es gab keine signifikanten Unterschiede zwischen den drei Versuchsgruppen: $\chi^2(2) = .654$, $p > .05$.

5.2. Glaubwürdigkeitskriterien

Die Anzahl an Aussagen, die codiert wurden, sehen wir folgt aus:

a) Eigenschaften des Experiments: 197

b) Eigenschaften von Autoren: 147

c) Eigenschaften der Daten: 123

d) Prüfen/Abgleichen: 78.

In Abbildung 1 sind die Verteilungen der Häufigkeiten der vier Codes für die drei Versuchsgruppen getrennt dargestellt. Für keinen der Codes bestehen signifikante Unterschiede zwischen den Versuchsgruppen. Die Werte des Kruskal-Wallis-Tests sind in Tabelle 1 aufgelistet.

Code	H	p	df
Eigenschaften des Experiments	0.993	>.05	2
Eigenschaften von Autoren	2.002	>.05	2
Eigenschaften der Daten	2.756	>.05	2
Prüfen/Abgleichen	2.996	>.05	2

Tabelle 1: Ergebnis des Kruskal-Wallis-Tests für die Unterscheidung der Versuchsgruppen für Häufigkeiten der Codes.

Die fünf häufigsten Subcodes und die entsprechende Anzahl an Codierungen sind:

a) Eig. des Experiments/Durchführung: 65

b) Eigenschaften der Daten/Streuung: 61

c) Eigenschaften von Autoren/Fehlbarkeit: 46

d) Prüfen/Abgleichen/Abgleich m. eig. Daten: 36

e) Eig. des Experiments/Menschentoleranz: 35

Die Verteilungen der Häufigkeiten für die einzelnen Subcodes sind in Abbildung 2 getrennt für die drei Versuchsgruppen dargestellt. Auf den ersten Blick sind auch hier keine signifikanten Unterschiede zwischen den Versuchsgruppen erkennbar. Für vier Subcodes zeigt der Kruskal-Wallis-Test keine signifikanten Unterschiede zwischen den Gruppen (vgl. Tabelle 2).

Subcode	H	p	df
Eig. d. Exp/Durchführung	1.350	>.05	2
Eig. d. Exp./Menschentoleranz	0.880	>.05	2
Eig. v. Aut./Fehlbarkeit	1.067	>.05	2
Eig. d. Daten/Streuung	3.810	>.05	2
Prüfen/Abgleich m.e. Daten	11.723	<.05	2

Tabelle 2: Ergebnisse des Kruskal-Wallis-Tests für die Unterschiede zwischen den Versuchsgruppen für die Häufigkeiten der fünf am häufigsten codierten Subcodes.

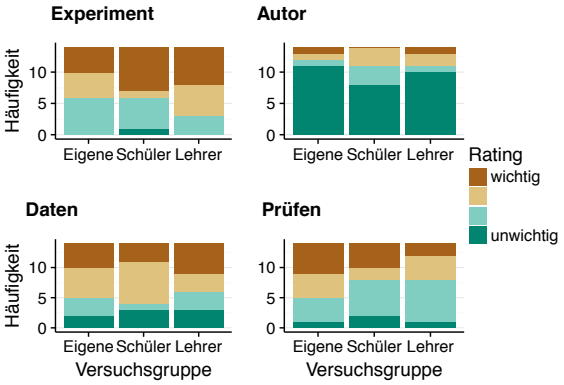

Abb. 3: Häufigkeiten der Einordnung der Codes in die Ränge dargestellt als Höhe der Teile der einzelnen Balken.

Für den Subcode „Abgleich mit eigenen Daten“ aus dem Code „Prüfen/Abgleichen“ ergibt sich ein signifikanter Unterschied zwischen den drei Versuchsgruppen. Ein multipler Vergleichstest nach Kruskal und Wallis (Field & Miles, 2012) ergibt, dass der signifikante Unterschied zwischen der Versuchsgruppe „Eigene Daten“ und „Lehrerdaten“ besteht. Zwischen den Versuchsgruppen „Eigene Daten“ und „Schülerdaten“ verfehlt der Test nur knapp den kritischen Wert. Hier ist also ein Unterschied zwischen den Versuchsgruppen mit Daten aus zweiter Hand und den Daten aus erster Hand zu erkennen.

5.3. Ranking der Wichtigkeit

Für das Ranking der Wichtigkeit der Codes durch die Probanden wurde die Einordnung der einzelnen Codes auf die verschiedenen Ränge gezählt (Abbildung 3). So ist es möglich zu vergleichen, wie oft ein Code z.B. als „am unwichtigsten“ oder auch „am wichtigsten“ gerankt wurde. Wichtig für das Verständnis von Abbildung 3 ist, dass ein höherer Anteil an grünen und dunkelgrünen Säulen für eine häufigere Einteilung in „eher unwichtig“ und „am unwichtigsten“ steht. So lässt sich nach Unterschieden in der Bewertung der einzelnen Codes suchen. Auch ein Vergleich des Rankings eines Codes durch verschiedene Versuchsgruppen ist möglich. Augenscheinlich wird der Code „Eigenschaften von Autoren“ deutlich häufiger als „unwichtig“ bewertet als andere Codes. Dies wird durch den Kruskal-Wallis-Test bestätigt: $H(3) = 42.512$, $p << .01$. Der multiple Vergleichstest nach Kruskal und Wallis zeigt, dass paarweise signifikante Unterschiede zwischen den drei Codes „Eigenschaften des Experiments“, „Eigenschaften der Daten“, „Prüfen/Abgleichen“ und dem Code „Eigenschaften von Autoren“ bestehen. Zwischen den anderen Paarungen dieser Codes bestehen keine signifikanten Unterschiede. Innerhalb der Codes bestehen bezüglich der Versuchsgruppen keine signifikanten Unterschiede (vgl. die Ergebnisse des Kruskal-Wallis-Tests in Tabelle 3).

Code	H	p	df
Eigenschaften des Experiments	1.067	>.05	2
Eigenschaften von Autoren	0.988	>.05	2
Eigenschaften der Daten	0.026	>.05	2
Prüfen/Abgleichen	1.592	>.05	2

Tabelle 3: Die Ergebnisse des Kruskal-Wallis-Tests für die Unterschiede zwischen Versuchsgruppen in den einzelnen Codes.

6. Diskussion

Die Ergebnisse der vorgestellten Studie lassen keine Unterschiede bezüglich der Versuchsgruppen erkennen. Offenbar stellten die unterschiedlichen Datensätze, die als Grundlage für die Interviews dienten, keinen Anlass dar, andere Kriterien für die Bewertung der Glaubwürdigkeit anzuwenden. Einen signifikanten Unterschied fanden wir zwischen den Versuchsgruppen „Eigene Daten“ und „Lehrerdaten“ beim Subcode „Prüfen/Abgleichen/Abgleich mit eigenen Daten“. Da zwischen „Eigenen Daten“ und den „Schülerdaten“ ein signifikanter Unterschied nur knapp verfehlt wurde, wird angenommen, dass hier eine relevante Unterscheidung zwischen Daten aus erster und Daten aus zweiter Hand ersichtlich ist. Diese Unterscheidung ist allerdings trivial und war so zu erwarten. Der Subcode „Abgleich mit eigenen Daten“ beschreibt Aussagen, die thematisieren, dass Daten mehr oder weniger glaubwürdig werden können, wenn man sie mit eigenen Daten vergleicht. In Abbildung 2 in der Spalte 42 für die Versuchsgruppe „Eigene Daten“ (E) ist zu erkennen, dass in nur zwei Interviews jeweils eine Aussage zu diesem Code gefunden wurde. Das ist gut erklärbar, denn für Schülerinnen und Schüler, die mit ihren eigenen Daten gearbeitet haben, bestand kein Anlass diese ihnen vorliegenden Daten mit ihren eigenen Daten zu vergleichen.

Die Ergebnisse liefern somit in ihrer Gesamtheit keine Hinweise darauf, dass die Verwendung von Daten aus erster und zweiter Hand einen Einfluss auf die Nutzung von Kriterien für die Bewertung von Glaubwürdigkeit hat. Dabei ist jedoch das Studiendesign zu beachten. Im Gegensatz zu anderen Studien (Hug & McNeill, 2008; Ludwig & Priemer, 2013; Magnusson & Palincsar, 2001) haben die Probanden dieser Untersuchung ein Realexperiment durchgeführt und eigene Daten aufgezeichnet. Dabei wurde darauf geachtet, dass die Ergebnisse der Probanden den gleichen Umfang und die gleiche Qualität haben, wie die fremden Daten. Es ist jedoch denkbar, dass die verschiedenen Datenquellen an Bedeutung gewinnen, wenn die Beteiligung am Experiment und damit das Wissen über das Experiment verändert wird.

Da es sich in diesem Versuch um eine Kombination von physikalischen Größen handelt, die in erster Näherung keinen Zusammenhang aufweisen, waren Schwierigkeiten beim Umgang mit Messunsicherheiten zu erwarten (Kanari & Millar, 2004). Das Design der Experimentierumgebung hat aber offenbar dazu geführt, dass die Daten auch den nicht vorhandenen Zusammenhang deutlich machen konnten. Dazu kommt, dass den Schülerinnen und Schülern die Suche im Hypothesenraum (Klahr, 2002) erspart blieb. Mögliche Hypothesen standen zur Auswahl. Dadurch war die Möglichkeit, dass kein Zusammenhang besteht, bereits vorgegeben. Werden Schülerinnen und Schüler aufgefordert eigene Hypothesen zu formulieren, wird diese Möglichkeit oft nicht beachtet (Kanari & Millar, 2004). Zudem thematisierten ein großer Teil der Aussagen der Schülerinnen und Schüler Messunsicherheiten oder damit verwandte Themen. Die Subcodes „Durchführung“, „Menschentoleranz“, „Fehlbarkeit“ und „Streuung“ thematisieren allesamt Fehlerquellen des Experiments, die Rückwirkung zwischen Experimentator und Experiment oder Abweichungen in den Daten. Somit wurden in vier der fünf am häufigsten genannten Subcodes Aspekte von Messunsicherheiten (Hellwig, 2012) angesprochen.

Bemerkenswert ist das Ergebnis, dass die Schülerinnen und Schüler zwar sehr oft über den Autor geredet haben, den Code „Eigenschaften von Autoren“ aber sehr häufig als „am unwichtigsten“ für die Bewertung der Glaubwürdigkeit bewertet haben.

Die Analyse der Häufigkeiten von Codierungen wurde zunächst aus der Annahme heraus durchgeführt, dass eine größere Anzahl an Aussagen, die mit einem Code codiert werden, ein Indiz für die Relevanz dessen ist, was mit dem Code umschrieben wird. Mehr Aussagen im Code „Eigenschaften von Autoren“ wurden also vorerst als Indiz dafür gewertet, dass die Probanden diesen Aspekt der Daten als relevant empfinden, wenn die Glaubwürdigkeit der Daten bewertet werden soll. Vielleicht haben die Schülerinnen und Schüler den ersten Interviewteil aber eher als Aufforderung zur Aufzählung von möglichst vielen Kriterien zur Glaubwürdigkeitsbewertung verstanden und haben dadurch ohne Rücksicht auf die Relevanz einzelner Punkte alles genannt, was ihnen in den Sinn kam. Demnach wäre die Häufigkeit ein ungenügendes Maß, um einen Rückschluss darauf zu gewinnen, welcher Aspekt im Denken der Probanden wichtiger ist als ein anderer.

Wie in anderen Studien zuvor, konnte das Fadenpendel auch hier sehr erfolgreich genutzt werden, um nicht-erwartungskonforme Daten zu erzeugen. Erfreulich hoch war die Anzahl von Schülerinnen und Schülern, welche nach der Durchführung des Experiments und der Konfrontation mit den Daten bereit waren, ihre eingangs aufgestellte, fachlich falsche Hypothese zugunsten der fachlich richtigen Hypothese aufzugeben. Es sei aber bemerkt, dass wir die Antworten der Schülerinnen und Schüler nur in Wechsel oder nicht Wechsel eingeteilt haben. Nicht alle Schülerinnen und Schüler waren vollständig von ihrem Ergebnis überzeugt. Vielmehr ergibt sich eine Abstufung an Reaktionen auf die Daten, ähnlich dem was bereits an möglichen Reaktionsty-

pen auf anomale Daten vorgeschlagen wurde (Chinn & Brewer, 1998).

7. Ausblick

Wie oben beschrieben finden sich zwar keine Hinweise darauf, dass der Datentyp einen Einfluss auf die Verwendung von Kriterien zur Glaubwürdigkeitsbewertung hat. Es muss aber das Studiendesign mitgedacht werden. Zwei Möglichkeiten für die Deutung dieses Ergebnisses könnten für weitere Forschungsarbeiten von Bedeutung sein. Es wurde vermutet, dass das Maß an Beteiligung an der Datenerhebung einen Einfluss auf die Wahrnehmung der Daten als „eigene Daten" hat. Die Beteiligung an der Datenerhebung vermittelt Wissen über die Daten, welches genutzt wird, wenn sie evaluiert werden. Fehlt dieses Wissen könnten andere Kriterien wieder an Bedeutung gewinnen. Es steht also die Frage im Raum, ob die Unterscheidung des Datentyps im Allgemeinen keine Rolle für die Verwendung von Glaubwürdigkeitskriterien spielt, oder ob der Datentyp diese Verwendung beeinflusst, wenn sich andere Rahmenbedingungen der Wahrnehmung der Daten verändern.

Die Codierung des Interviewmaterials hatte notwendigerweise einen Informationsverlust zur Folge. Verschiedene Aussagen, die zwar thematisch gruppiert werden können, unterscheiden sich aber dennoch in Details. Diese Informationen können für das Verständnis der Vorstellungen der Schülerinnen und Schüler von Bedeutung sein. Eine inhaltliche Analyse der Aussagen für die fünf häufigsten Subcodes wird deshalb angestrebt.

8. Literatur

[1] Bentele, G., Brosius, H.-B., & Jarren, O. (2012). *Lexikon Kommunikations- und Medienwissenschaft*. Springer-Verlag.

[2] Bortz, P. D. J., & Schuster, P. D. C. (2010). Analyse von Häufigkeiten. In Statistik für Human- und Sozialwissenschaftler (S. 137–152). Springer Berlin Heidelberg.

[3] Chinn, C. A., & Brewer, W. F. (1993). The Role of Anomalous Data in Knowledge Acquisition: A Theoretical Framework and Implications for Science Instruction. Review of Educational Research, 63(1), 1–49.

[4] Chinn, C. A., & Brewer, W. F. (2001). Models of Data: A Theory of How People Evaluate Data. *Cognition and Instruction*, *19*(3), 323–393.

[5] Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological Bulletin, 70(4), 213.

[6] Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science education*, *84*(3), 287–312.

[7] Field, A., & Miles, J. (2012). *Discovering Statistics Using R*. London ; Thousand Oaks, Calif: Sage Publications Ltd.

[8] Gwet, K. L. (2012). *Handbook of Inter-Rater Reliability* (3. Aufl.). Advanced Analytics, LLC.

[9] Heinicke, S. (2012). *Aus Fehlern wird man klug*. Logos Verlag, Berlin.

[10] Hellwig, J. (2012). *Messunsicherheiten verstehen* (Dissertation). Ruhr-Universität Bochum.

[11] Hug, B., & McNeill, K. L. (2008). Use of First-hand and Second-hand Data in Science: Does data type influence classroom conversations? International Journal of Science Education, 30(13), 1725–1751.

[12] Kanari, Z., & Millar, R. (2004). Reasoning from data: How students collect and interpret data in science investigations. *Journal of Research in Science Teaching*, *41*(7), 748–769.

[13] Klahr, D. (2002). *Exploring Science – The Cognition and Development of Discovery Processes*. The MIT Press.

[14] Koslowski, B. (1996). *Theory and Evidence: The Development of Scientific Reasoning*. MIT Press.

[15] Kuhn, D. (2002). What is Scientific Thinking and How Does it Develop? In U. Goswami (Hrsg.), *Blackwell Handbook of Childhood Cognitive Development* (1. Aufl., S. 497–523).

[16] Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. Biometrics, 33(1), 159–174.

[17] Ludwig, T., & Priemer, B. (2013). Secondary School Students' Reasoning from Anomalous Data. Gehalten auf der NARST Annual International Conference, Puerto Rico.

[18] Magnusson, S. J., & Palincsar, A. S. (2001). The interplay of first-hand and second-hand investigations to model and support the development of scientific knowledge and reasoning, in Cognition and instruction: twenty-five years of progress (S. 151–194), Psychology Press.

[19] Nicolaidou, I., Kyza, E. A., Terzian, F., Hadjichambis, A., & Kafouris, D. (2011). A framework for scaffolding students' assessment of the credibility of evidence. *Journal of Research in Science Teaching*, *48*(7), 711–744.

[20] Rieh, S. Y., & Danielson, D. R. (2007). Credibility: A multidisciplinary framework. *Annual Review of Information Science and Technology*, *41*(1), 307–364.

[21] Tesch, M. (2005). *Das Experiment im Physikunterricht*. (H. Niedderer, H. FIschler, & E. Sumfleth, Hrsg.). Logos Verlag.

[22] Wilson, P. (1983). *Second-Hand Knowledge - An Inquiry Into Cognitive Authority*. Greenwood Press.