Steffen Brockmüller[1]
Mathias Ropohl[1]

[1]Universität Duisburg-Essen

Ein Mixed-Methods-Zugang zur Untersuchung von Schwierigkeiten beim Auswerten von Versuchsdaten

Theoretischer Hintergrund

In den Naturwissenschaften sind Lernende beim Auswerten von Versuchsdaten in Denk- und Arbeitsweisen involviert, die (1) das Daten Aufbereiten und (2) Interpretieren, (3) das Schlussfolgern, (4) die kritische Reflexion des Versuchs sowie das (5) Finden möglicher Generalisierungen einschließen. Derartige Operationen fallen Lernenden häufig schwer, so bspw. das Ableiten von Trends aus Daten, das Belegen von Schlussfolgerungen mit Versuchsergebnissen oder das Berücksichtigen von Gütekriterien beim Beurteilen von Daten (Jeong, Songer & Lee, 2007; Lubben & Millar, 1996; Sandoval & Millwood, 2005).

Insbesondere für prozedural komplexere Versuche von Oberstufenlernenden der Chemie ist allerdings bislang wenig darüber bekannt, welche Schwierigkeiten bestehen. Auch ist nicht geklärt, inwiefern diese Schwierigkeiten mit dem Vorwissen der Schülerinnen und Schüler zusammenhängen. Entsprechend gängigen Konzeptualisierungen des *Scientific Reasonings* (Opitz et al., 2017) werden prozedurales (d.h. auf das Auswerten von Versuchsdaten bezogenes) Wissen, epistemisches (d.h. auf Konzepte und Begründungen wissenschaftlichen Arbeitens bezogenes) Wissen und Fachwissen als mögliche Bedingungsfaktoren für die Performanz von Lernenden beim Experimentieren allgemein und beim Auswerten der darin erhobenen Daten im speziellen vermutet (Osborne, 2018).

Fragestellung

In der hier beschriebenen Untersuchung werden diese als Schwierigkeiten salient werdenden situationsspezifischen Denkprozesse der Lernenden aus ihrem Kommunizieren und Handeln beim Auswerten von Experimenten abgeleitet und in Bezug zu relevanten Wissensdispositionen gesetzt (Blömeke, Gustafsson & Shavelson, 2015; Vorholzer & von Aufschnaiter, 2020). Geklärt werden sollen dadurch folgende Forschungsfragen:

1. Welche Schwierigkeiten bei der Datenauswertung bestehen bei Lernenden der Oberstufe beim Auswerten aus chemischen Versuchen gewonnener Daten?
2. Welche Zusammenhänge bestehen zwischen dem Auftreten dieser Schwierigkeiten und dem Fachwissen sowie dem prozeduralen und epistemischen Wissen der Schülerinnen und Schüler?

Methode

Im Rahmen dieses Vorhabens läuft gegenwärtig (Stand Oktober 2021) eine Studie, bei der Lernende der Q1 und Q2 ($N = 100$) beim Durchführen und Auswerten von zwei inhaltlich und strukturell ähnlichen Versuchen in Zweiergruppen videographiert werden. Bei den Versuchen handelt es sich um hypothesenprüfende Versuche zu Säuren und Basen. Im ersten Setting untersuchen die Lernenden zwei Salzlösungen auf ihre Pufferwirkung. Im zweiten Setting vergleichen sie die Säurestärke zweier Säurelösungen. Die Problemstellung, Forschungsfrage, Hypothese und Durchführung der Versuche werden den Lernenden vorgegeben. Die Daten

nehmen die Lernenden mithilfe digitaler Sensoren auf, die eine Darstellung von abgelaufener Zeit, zugetropftem Volumen von Säure bzw. Base sowie pH-Wert in Form separater Tabellen ermöglichen. Ein Begleitheft mit offenen Fragen, die auf die o.g. Operationen des Auswertens von Versuchsdaten abheben, unterstützen die Lernenden beim Auswerten der Daten in diesen Tabellen.

Die so erhaltenen Videodaten werden, ergänzt durch die Lernendenantworten, mithilfe einer kategorienbildenden qualitativen Inhaltanalyse (Mayring, 2015) induktiv auf Schülerschwierigkeiten untersucht. Als Selektionskriterium dient dabei das Auftreten einer *unmittelbar beobachtbaren Schwierigkeit*, welches durch ein von Kechel (2016) adaptiertes Kodiermanual operationalisiert wird. Um das Vorwissen der Lernenden zu erheben, steht für die drei relevanten Wissensbereiche je ein Instrument zur Verfügung. Diese wurden aus bestehenden Skalen adaptiert und anschließend pilotiert (s. Brockmüller & Ropohl, 2021). Die Instrumente kommen vor der Durchführung der Versuche zum Einsatz und werden IRT-basiert ausgewertet. Zur Kontrolle der kognitiven Fähigkeiten wird eine nonverbale KFT-Skala verwendet.

Ergebnisse einer Pilotstudie

Im Zuge einer Erprobung des ersten Experimentiersettings (s. Brockmüller & Ropohl, 2021) wurden $N = 22$ Lernende beim Durchführen des Versuchs zur Untersuchung eines Säure-Base-Puffers in Zweiergruppen videographiert. Die Aufgabe für die Lernenden bestand zunächst darin, aus der technisch bedingt großen Anzahl an Datenpaaren systematisch aufzutragende Daten auszuwählen und in Form von Titrationskurven darzustellen. Darauf aufbauend strukturierten offene Aufgaben ihren Auswertungsprozess.

Die resultierenden elf Videos wurden mithilfe des oben beschriebenen Kodiermanuals inhaltsanalytisch ausgewertet. Dazu analysierten zwei Kodierer jeweils alle 11 Videos mithilfe von MAXQDA 2020 und erstellten dabei induktiv je ein Kategoriensystem von beobachteten Schwierigkeiten. Anschließend wurden diese beiden Systeme kritisch verglichen und zusammengeführt. Zur Prüfung der Intercoder-Reliabilität dieses Kategoriensystems wurde dieses anschließend erneut auf die 11 Videos angewendet. Beteiligt war ein Kodierer aus dem ersten Durchgang sowie ein dritter, neu an Kodiermanual und Kategoriensystem geschulter, Kodierer.

Ein Vergleich der Kodierungen dieses zweiten, deduktiven Durchgangs liefert eine nach Landis und Koch (1977) als substanziell zu bezeichnende Intercoder-Übereinstimmung von $\kappa = .62$, was angesichts der hohen Inferenz des zu kodierenden Merkmals ein adäquates Ergebnis darstellt. Das finale Kategoriensystem ist in Tabelle 1 einzusehen. Im Durchschnitt wurde eine Schwierigkeitskategorie für $M = 3.34$ Proband:innen kodiert ($SD = 4.33$). Pro Proband:in wurden durchschnittlich $M = 5.55$ Schwierigkeiten diagnostiziert ($SD = 1.60$, $Min = 3$, $Max = 9$). Es überwiegen bislang deutlich Kategorien für den Bereich *Daten aufbereiten*, was als Artefakt einer häufig unvollständigen Bearbeitung des Settings zu interpretieren ist. Eine Reduzierung des Zeitaufwands der Durchführung der Versuche sowie eine veränderte Schwerpunktsetzung der Auswertung schafft diesbezüglich für die Hauptstudie Abhilfe.

Tabelle 1
Inhaltsanalytisch abgeleitete Schwierigkeitskategorien

	#	Kurzbezeichnung der Kategorie
Daten aufbereiten	**1**	Unsicherheit beim Skalieren der Diagramme
	2	Auswahl von Variablen
	3	Zuweisung von Variablen auf Achsen
	4	Uneinheitl. Auflösung der Diagramme
	5	Uneinheitliche Intervallschritte innerhalb *eines* Diagramms
	6	Auflösung d. Diagramme zu groß/ klein
	7	Fehler beim Eintragen
	8	Eintragen ausgedachter Werte
	9	Auswahl v. Messdaten unsystematisch
	10	Keine Auswahl von Messwerten
	11	Zu grobe Auswahl von Messwerten
	12	Auswahl der Messwerte in Zeitintervallen
	13	Auswahl d. Messwerte durch Abzählen
	14	Auswahl v. Messwerten - Unsicherheit
	15	Nicht in *einem* Diagramm verglichen
	16	Unnötig Werte am Ende eingetragen
	F	Ungerichtete Frustration (Aufbereiten)
Daten interpretieren	**1**	Beschreibung d. Kurvenverlaufs ungenügend
	2	Interpretationsansatz ungenau
	3	Interpretation fachlich falsch
	4	Falsches Ablesen
	5	Mangelndes Vertrauen in eigenes Vorwissen
	F	Ungerichtete Frustration (Interpretieren)
Schlussf.	**1**	Kein Hypothesenbezug
	2	Keine Begründung der Schlussfolgerung
	3	Falsches Versuchsziel verfolgt
	F	Ungerichtete Frustration (Schlussfolgern)
Beurteilen	**1**	Fehler erkannt, aber nicht identifiziert
	2	Unsicherheit im Umgang mit anomalem Messwert
	3	Schweren Fehler nicht erkannt
	F	Ungerichtete Frustration (Beurteilen)
Ge	**1**	Folgeexperiment ungenau beschrieben
	F	Ungerichtete Frustration (Generalisieren)
Global	**1**	Ungültiges Abkürzen durch Vorwissen
	2	Zeit nicht genutzt
	3	Zeitmangel

Anmerkungen. Schlussf. = Schlussfolgern. Ge = Generalisieren.

Binär-logistische Regressionsanalysen des Einflusses der Personenfähigkeiten in den verschiedenen Wissensbereichen auf das Auftreten dieser Kategorien zeigen nicht zuletzt aufgrund der geringen Stichprobengröße in der Pilotstudie ein wenig konklusives Bild. So steigt die Wahrscheinlichkeit für unsystematische Auswahl von Messwerten zur Aufbereitung in Graphen (Schwierigkeit 9 beim *Aufbereiten*) mit steigendem epistemischen Wissen (B = 1.25, SE = .61, Wald = 4.16, p =.042, exp(B) = 3.47, CI[1.05, 11.49]). Die Wahrscheinlichkeit, mehrere zu vergleichende Graphen nicht einheitlich zu skalieren (Schwierigkeit 4 beim *Aufbereiten*), sinkt mit steigendem Fachwissen (B = -2.39, SE = 1.12, Wald = 4.53, p =.033, exp(B) = 0.092, CI[.01,.83]). In beiden Fällen lassen große Konfidenzintervalle allerdings auf die Notwendigkeit der geplanten Stichprobenvergrößerung für derartige Analysen in der Hauptstudie schließen. Weitere Regressionskoeffizienten sind nicht signifikant bzw. wurden wegen einer geringen Zahl an Fällen nicht berechnet.

Literatur

Blömeke, S., Gustafsson, J.-E., & Shavelson, R. J. (2015). Beyond dichotomies: Competence viewed as a continuum. *Zeitschrift für Psychologie*, 223 (1), 3-13.

Jeong, H., Songer, N.B. & Lee S.-Y. (2007). Evidentiary Competence: Sixth Graders' Understanding for Gathering and Interpreting Evidence in Scientific Investigations. *Research in Science Education* 37 (1), 75-97.

Kechel, J.-H. (2016). *Schülerschwierigkeiten beim eigenständigen Experimentieren. Eine qualitative Studie am Beispiel einer Experimentieraufgabe zum Hooke'schen Gesetz*. Berlin: Logos.

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159-174.

Lubben, F. & Millar, R. (1996). Children's ideas about the reliability of experimental data. *International Journal of Science Education*, 18(8), 955-968.

Mayring, P. (2015). *Qualitative Inhaltsanalyse. Grundlagen und Techniken* (12., überarb. Aufl.). Weinheim: Beltz.

Opitz, A., Heene, M. & Fischer, F. (2017). Measuring scientific reasoning – a review of test instruments. *Educational Research and Evaluation*, 23 (3-4), 78-101.

Osborne, J. (2018). Styles of scientific reasoning: What can we learn from looking at the product, not the process, of scientific reasoning? In Fischer, F., Chinn, C. A., Engelmann, K., & Osborne, J. (eds.), *Scientific reasoning and argumentation: The roles of domain-specific and domain-general knowledge*. London: Routledge.

Sandoval, W.A. & Millwood, K.A. (2005). The Quality of Students' Use of Evidence in Written Scientific Explanations, *Cognition and Instruction*, 23 (1), 23-55.

Vorholzer, A. & Aufschnaiter, C. v. (2020). Dimensionen und Ausprägungen fachinhaltlicher Kompetenz in den Naturwissenschaften – ein Systematisierungsversuch. *Zeitschrift für Didaktik der Naturwissenschaften*, 26, 1-18.