

Guido Haag
 Jochen Scheid
 Patrick Löffler
 Alexander Kauertz

Universität Koblenz-Landau

Entwicklung eines Instruments zur Beurteilung manueller Kompetenzen

Abstract: Experimente sollen dazu dienen, Schülerinnen und Schülern die Grundprinzipien von SI und NoS näher zu bringen. Schülerinnen und Schüler sollen dabei erlernen, wie wissenschaftliche Kenntnisse erworben werden und was wissenschaftliche Methoden charakterisiert. Weitere Ziele sind mit großen Erwartungen verknüpft, so z. B. einer Verbesserung der experimentellen Kompetenzen oder der Motivation (Kircher, Girwidz & Häußler, 2015). Das Experiment nimmt daher eine zentrale Stellung im naturwissenschaftlichen Unterricht ein (vgl. z. B. Bybee, R., 2000; Hofstein, A., 2004). Die Entwicklung experimenteller Kompetenzen ist daher ein wesentliches Ziel des Physikunterrichts (Schecker, Neumann, Theyßen, Eickhorst, & Dickmann, 2016). Dies spiegelt sich auch international in Richtlinien wider (OECD, 2007; KMK, 2005). Leider werden diese Erwartungen oft nicht erfüllt (Harlen, 1999; Hofstein & Lunetta, 2004). Um die Gründe für das Ausbleiben des erhofften Lernzuwachses zu erforschen setzte die bisherige Forschung zu experimentellen Kompetenzen meist auf Paper- und Pencil-Tests (Parchmann, Schecker 2006; Schecker, Theyßen und Schreiber 2009). Die Validität solcher Untersuchung zur Beurteilung von Kompetenzen der manuellen Ausführung wird allerdings angezweifelt (Parchmann, Schecker 2006; Schecker, Theyßen und Schreiber 2009; Schecker et al., 2016), denn es zeigen sich in empirischen Studien nur geringe Zusammenhänge von Leistungsmessungen in schriftlichen Tests und Realsituationen (Gut-Glanzmann, C., 2012; Schreiber, Theyßen & Schecker, 2014). Um diese Lücke zu schließen, stellen wir ein Kategoriensystem auf, mit dem wir Experimente vergleichbar machen wollen um experimentelle Kompetenzen in verschiedenen Lernumgebungen beurteilen zu können. Novizen zeigen schlechtere experimentelle Kompetenzen als Experten. Kompetenzen sind durch Fehler abbildbar. Während der Durchführung treten Fehler der manuellen Ausführung auf. Wir erstellen ein Kategoriensystem für diese Fehler.

Keywords: Naturwissenschaftliche Bildung, *Experimentieren, Kompetenzbeurteilung*

Theoretischer Hintergrund und Desiderat: Neben äußeren Ursachen wie Rahmenbedingungen (Kirchner, 2015), Zielkonflikten (Hammann, Phan, Ehmer & Bayrhuber, 2006) oder Über- bzw. Unterforderung, die sich aus zu offenen bzw. zu detaillierten Experimentieranweisungen ergeben (Börlin, 2012) werden auch als individuelle Ursachen fehlende Routine (Kircher et al., 2015), mangelnde Beherrschung der Messgeräte (Hopf, 2007) sowie generelle Überforderung durch die Vielzahl der Anforderungen (vgl. z. B. Börlin, 2012) genannt. Daher ist die Frage, wie sich die Fähigkeit des Experimentierens entwickelt und wie sie im Unterricht gefördert werden kann, ein aktuelles Forschungsfeld (Labudde und Möller, 2012). Experimente gliedern sich in Phasen, die z. B. als Planung, Durchführung und Auswertung bezeichnet werden (Schecker et al., 2016). In unterschiedlichen Phasen zeigen sich unterschiedliche Kompetenzen (Börlin, 2012). Maiseyenko, Nawrath und Schecker haben 2013 dazu ein Modell mit sieben Dimensionen experimenteller Kompetenz erarbeitet: 1) Fragestellung entwickeln, 2) Vermutung / Hypothese aufstellen, 3) Experiment planen, 4) Versuch funktionsfähig aufbauen, 5) Beobachten / Messen / Dokumentieren, 6) Daten aufbereiten, 7) Schlüsse ziehen / diskutieren. Im Rahmen dieses Modells lassen sich die Kompetenzen *Versuch aufbauen, Beobachten, Messen* und *Dokumentieren* der Durchführungsphase des Experiments

zuordnen. Bisher wurden überwiegend diejenigen Aspekte experimenteller Kompetenz untersucht, die sich gut mit Paper- und Pencil-Tests erfassen lassen. Dazu gehören vor allem Kompetenzen, die sich in den Phasen der Planung und der Auswertung des Experiments zeigen. Als Grund hierfür wird genannt, dass Tests in Realsituationen in Durchführung und Auswertung aufwendig sind und sich somit nicht für Large-Scale-Untersuchungen anbieten (Schecker et al., 2016). Harlen (2010) führt dazu aus, dass mit Hilfe schriftlicher Tests zur Beurteilung experimenteller Kompetenz zwar schnell ein breiter Bereich themengebundener Kompetenzen abgefragt werden kann, stellt allerdings die Validität der schriftlichen Abfrage von Handlungskompetenzen in Frage. Auch Parchmann, Schecker (2006) sowie Schecker, Theyßen und Schreiber (2009) bezweifeln die Validität schriftlicher Tests zur Beurteilung manueller Kompetenzen der Durchführungsphase. Ausführungsfehler sind im allgemeinen kompetenzbasiert. Die Abgrenzung von Fehlertypen, ihre Häufigkeit und Auswirkungen erlaubt eine Einschätzung, wo Prävention und Therapie nötig ist (Müller, 2003). Es besteht daher Anlass zur Untersuchung explizit manueller experimenteller Kompetenzen anhand beobachteter Fehler während der Durchführungsphase von Experimenten.

Forschungsfragen und Hypothesen:

In Experimenten werden unter festgelegten und kontrollierten Rahmenbedingungen Messungen an Objekten durchgeführt (Hartinger, 2014). In quantitativen Experimenten kommen dabei Messinstrumente zum Einsatz um Werte zu bestimmen (Kircher et al., 2015, p. 234). Möchte man experimentelle Kompetenzen anhand von Sichtstrukturmerkmalen untersuchen, so ist unter anderem die sichere Identifikation der Phase eine Grundvoraussetzung für eine hoch inferente Beurteilung experimentellen Handelns (Börlin, 2012, p. 49). Wir untersuchen *Handlungen* am Experiment. Darunter verstehen wir in Anlehnung an Harlen (2010): Schritte mit dem Ziel der Beschreibung der für die Fragestellung relevanten Größen. Es gibt daher eine Abfolge *notwendiger Schritte*. Unter notwendigen Schritten verstehen wir alle Handlungen, die den kompletten experimentellen Durchlauf ermöglichen. Die Menge der Notwendigen Schritte lässt sich in zwei Subkategorien aufteilen: a) *Unwirksame Schritte*: *Notwendige Schritte*, die keinen Einfluss auf den Wert des zu Ende geführten Experiments nehmen können b) *Wirksame Schritte*: *Notwendige Schritte*, die Einfluss auf den Wert des zu Ende geführten Experiments nehmen können. Wir unterteilen wir die Durchführungsphase in wirksame Schritte. Experimente unterscheiden sich damit voneinander, sie besitzen *Attribute*. Um Experimente bezüglich der Kompetenzen der manuellen Ausführung untereinander vergleichbar zu machen entwickeln wir ein Konstrukt anhand dieser Attribute. Dazu betrachten wir die Anzahl der gemessenen Größen, die Anzahl der notwendigen Messungen, die Anzahl der eingesetzten Arten von Messinstrumenten und der Anzahl der wirksamen Schritte.

F1: Wie lassen sich Experimente anhand ihrer Attribute vergleichbar machen?

H1: Für jedes Experiment lässt sich die Anzahl der gemessenen Größen, der Messungen, der eingesetzten Arten von Messinstrumenten und der wirksamen Schritte eindeutig bestimmen.

Abweichungen vom erwarteten Handlungsziel und die ungeplant zu einer Veränderung des Ergebnisses führen bezeichnen wir als *Fehler*. Fehler treten bei der quantitativen Variablenkontrolle auf (Kirchner, 2013, p. 26), beim Umgang mit Messgeräten (Schenk et al., 2014, p. 12), der Auswahl von Quantitäten und Qualitäten (Millar, Lubben et al., 1994) oder durch Nichtbefolgen der Anleitung (Börlin, 2012). Wir bezeichnen diese Fehler als *Variablenkontrollfehler*, *Messgerätefehler*, *Relativer Fehler* und *Abweichungsfehler*.

F2: Welche Fehlerkategorien in der Durchführungsphase von Experimenten lassen sich identifizieren und wie häufig treten sie auf?

H2: Die genannten Fehlerkategorien sind mit Hilfe eines Beurteilertrainings empirisch nachweisbar.

F3: Wie eignen sich manuelle *Fehler* zur Beschreibung experimenteller Kompetenzen?

H3: Schüler mit höheren Kompetenzen erzielen einen geringeren *Fehlerquotienten* Anzahl *Fehler*/Anzahl der durchlaufenen wirksamen Schritte.

Vorstudie: Es wurden sechs Experimente ausgewählt und Experimentieranleitungen dazu verfasst. Nach einem Training beurteilten vier Experten mit Hilfe eines weiteren Manuals die Attribute der Versuche. Dabei wurde von den Ratern nicht nur die jeweilige Anzahl genannt, zusätzlich mussten ausformulieren um einen eindeutigen Vergleich sicher zu stellen. Es wurde ein Beobachtungsbogen erstellt, in dem die im Expertenrating gefundenen wirksamen Schritte sowie die theoretisch hergeleiteten Fehlerkategorien aufgelistet sind. Aus Gründen der Redundanz wurde die Kategorie *Sonstige Fehler* hinzugefügt. Zudem wird erfasst, wer welchen Schritt wie oft vollzogen hat. Im Anschluss durchliefen 26 Schülerinnen und Schüler der 8. und 9. Klasse einer Gemeinschaftsschule das Setting in Partnerarbeit, wobei sie videografiert wurden. Nach einem Training beurteilten 3 Experten die Vignetten und dokumentierten Fehler und Anzahl der jeweils durchlaufenen wirksamen Schritte.

Ergebnisse: Zur Beurteilung der Reliabilität des Attributeratings wurde die Intraklassenkorrelation (ICC) ermittelt, da es sich um intervallskalierte Werte handelt (Wirtz und Caspar, 2002). Es ergaben sich folgende Werte für Cronbachs α : Anzahl gemessener Größen: .963, Anzahl der notwendigen Messungen: .978, Anzahl der Arten von Messinstrumenten: 1.000, Anzahl der wirksamen Schritte: .770.

Bei der Beurteilung der Reliabilität des Fehlerratings handelt es sich auch um intervallskalierte Werte. Da es sich aber meist um eine 1 Fehler/0 Fehler-Beurteilung handelt, entsteht keine Varianz. Damit liefert eine ICC keine bedeutsame Reliabilität (Wirtz und Caspar, 2002). Zur Beurteilung der Reliabilität des Fehlerratings wurde deshalb die prozentuale Übereinstimmung gefundener Fehler herangezogen. Sie ist aussagekräftig, da Zufallsübereinstimmungen eher unwahrscheinlich sind (Wirtz und Caspar, 2002). Den etwa 80 Fehlerfunden standen etwa 1400 Missings gegenüber. Eine Zufallsbereinigung liefert damit nur geringfügig kleinere Werte. Im Vergleich aller möglichen Raterpaarungen ergaben sich folgende Werte: Rater B / Rater H: 58%, Rater B / Rater S: 62%, Rater H / Rater S: 71%.

Diskussion und Ausblick: Die Werte der prozentualen Übereinstimmung sind nicht befriedigend. Zum Teil erklären sich die Unterschiede dadurch, dass zwar gemeinsam Fehler gefunden wurden, diese jedoch in andere Kategorien eingruppiert wurden. Dadurch entstehen zwei Fälle, in denen keine Übereinstimmung vorliegt, obwohl der Fehler von allen gefunden wurde. Betrachtet man nur den Fehlerfund, ohne Kategorie, so ergeben sich höhere prozentuale Übereinstimmungen. An dieser Stelle muss das Manual ausgeschärft werden.

Wir untersuchen manuelle experimentelle Kompetenz als individuelles Personenmerkmal. In Partnersettings treten dabei weitere, schwer kontrollierbare Effekte auf. Es ist valider und reliabler, in der Hauptstudie Einzelexperimente zu untersuchen, auch wenn dies mit einem erhöhten Studienaufwand verbunden ist.

In der Hauptstudie soll die experimentelle Kompetenz der manuellen Durchführung durch einen Quotienten Fehler / wirksamer Schritt abgebildet werden. Dabei soll die Validität des Messinstruments durch weitere Tests bestätigt werden. Wir erwarten Korrelationen zwischen diesem Quotienten und den weiteren Attributen von Versuchen.

Literatur

- Börlin, J. (2012). *Das Experiment als Lerngelegenheit: Vom interkulturellen Vergleich des Physikunterrichts zu Merkmalen seiner Qualität*. Berlin.
- Bybee, R. (2000). Teaching Science as Inquiry. In J. A. Minstrell & American Association for the Advancement of Science. (Eds.), *Inquiring into inquiry : learning and teaching in science* (pp. 20–47). Washington D.C: American Association for the Advancement of Science.
- Gut-Glanzmann, C. (2012). *Modellierung und Messung experimenteller Kompetenz: Analyse eines large-scale Experimentiertests*. Zugl.: Basel, Univ., Diss., 2012. *Studien zum Physik- und Chemielernen: Vol. 134*. Berlin: Logos-Verlag
- Hammann, M., Phan, T. H., & Bayrhuber, H. (2008). Experimentieren als Problemlösen: Lässt sich das SDDS-Modell nutzen, um unterschiedliche Dimensionen beim Experimentieren zu messen? In M. Prenzel, I. Gogolin, & H.-H. Krüger (Eds.), *Zeitschrift für Erziehungswissenschaft Sonderheft: Vol. 8*.
- Harlen, W. (1999). *Effective teaching of science: A review of research. SCRE publication Using research series: Vol. 21*. Glasgow: Scottish Council for Research in Education.
- Hartinger, A. (2014). Experimente und Versuche. In D. von Reeken (Ed.), *Dimensionen des Sachunterrichts: Vol. 3. Handbuch Methoden im Sachunterricht* (3rd ed., pp. 68–75). Baltmannsweiler: Schneider-Verl. Hohengehren.
- Hodson, D. (1993). Re-thinking Old Ways: Towards A More Critical Approach To Practical Work In School Science. *Studies in Science Education*, 22 (1), 85–142.
- Hofstein, A., & Lunetta, V. N. (2004). The laboratory in science education: Foundations for the twenty-first century. *Science Education*, 88(1), 28–54.
- Hofstein, A., & Mamlok-Naaman, R. (2007). The laboratory in science education: the state of the art. *Chemistry Education Research and Practice*, 8(2), 105–107.
- Müller, A. (2003). Fehlertypen und Fehlerquellen beim Physiklernen – Was weiß die Denkpsychologie? In: PdN–Physik, 1/52; Jahrgang 2003
- Hopf, M. (2007). *Problemorientierte Schülerexperimente*. Zugl.: München, Univ., Diss., 2007. *Studien zum Physik- und Chemielernen: Vol. 68*. Berlin: Logos-Verlag.
- Kircher, E., Girwidz, R., & Häußler, P. (2015). *Physikdidaktik*. Berlin, Heidelberg: Springer Berlin Heidelberg.
- Kirchner, S. (2013). Der Umgang mit Variablen bei offenen Experimentieraufgaben im Physikunterricht: Eine Beobachtungsstudie am Beispiel der Konstruktion von auftriebserzeugenden Profilen für ein Windradmodell (Dissertation). Humboldt-Universität zu Berlin, Berlin.
- KMK, (2005). *Bildungsstandards im Fach Physik für den Mittleren Schulabschluss: Beschluss vom 16.12.2004*. München: Wolters Kluwer.
- Labudde, P., & Möller, K. (2012). Stichwort: Naturwissenschaftlicher Unterricht. *Zeitschrift für Erziehungswissenschaft*, 15(1). Retrieved from 10.1007/s11618-012-0257-0
- Maiseyenko, V., Nawrath, D. & Schecker, H. (2011). Modellbasierte Förderung und Diagnose von Experimentierkompetenz. In: D. Höttecke (Hg.): *Naturwissenschaftliche Bildung als Beitrag zur Gestaltung partizipativer Demokratie*. Gesellschaft für Didaktik der Chemie und Physik. Jahrestagung in Potsdam 2010. Münster: Lit Verlag.
- Papula, L. (2015). *Mathematik für Ingenieure und Naturwissenschaftler - Anwendungsbeispiele*. Wiesbaden: Springer Fachmedien Wiesbaden.
- Parchmann, I., & Schecker, H. (2006). Modellierung naturwissenschaftlicher Kompetenz. *Zeitschrift für Didaktik der Naturwissenschaften*, 12.
- Schecker, H., Neumann, K., Theyßen, H., Eickhorst, B., & Dickmann, M. (2016). Stufen experimenteller Kompetenz. *Zeitschrift für Didaktik der Naturwissenschaften*, 22 (1), 197–213.
- Schecker, H., Theyßen, H., & Schreiber, N. (2009). Experimentelle Kompetenz messen?! *Physik und Didaktik in Schule und Hochschule*, 3(8), 92–101.
- Schenk, W., Kremer, F., Beddies, G., Franke, T., Galvosas, P., & Rieger, P. (2014). *Physikalisches Praktikum*. Wiesbaden: Springer Fachmedien Wiesbaden.
- Schreiber, N., Theyßen, H., & Schecker, H. (2014). Diagnostik experimenteller Kompetenz: Kann man Realexperimente durch Simulationen ersetzen? *Zeitschrift für Didaktik der Naturwissenschaften*, 20 (1).
- Tesch, M., Duit, R. (2002). Zur Rolle des Experiments im Physikanfangsunterricht. In: V. Nordmaier (Hg.): *Didaktik der Physik: Beiträge zur Frühjahrstagung der DPG –Leipzig*
- Wagner, J., Tipler, P. A., Mosca, G., Basler, M., & Kommer, C. (Eds.). (2015). *Lehrbuch. Physik für Wissenschaftler und Ingenieure: [der Begleiter bis zum Bachelor]* (7. dt. Aufl.). Berlin: Springer Spektrum.
- Wirtz, M., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität: Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen*. Göttingen: Hogrefe Verl. für Psychologie.