Johannes Schulz Humboldt-Universität zu Berlin
Burkhard Priemer

Messung von Kompetenzen im Umgang mit Messunsicherheiten

Die Beurteilung der Qualität von Messungen ist essentiell in Naturwissenschaft und Technik. Über die Fachdomänen hinaus ist die Behandlung von Messunsicherheiten aber auch in der Lehre bedeutsam, denn auch hier werden Folgerungen aus empirischer Evidenz gezogen. Um in diesem Bereich Kompetenzen von Schüler*innen valide und reliabel zu erfassen, bedarf es eines geeigneten Testinstrumentes. In diesem Beitrag werden die Ergebnisse von empirischen Erhebungen zur Entwicklung eines solchen Testinstrumentes mit Schüler*innen für zwei Teilbereiche „Verlässlichkeit der Messung“ und „Vergleich von Messwerten“ des Themenfeldes Messunsicherheiten vorgestellt.

Theoretischer Hintergrund

Daten aus Experimenten lassen sich ohne Kompetenzen im Umgang mit Messunsicherheiten wissenschaftlich nicht sinnvoll bewerten. Trotzdem wird das Thema Messunsicherheiten in der Schule oft vernachlässigt. Um zukünftige Studien in diesem Bereich durchzuführen, bedarf es eines passenden Testinstrumentes. Auf Basis eines Sachstrukturmodells und einer Reduktion für die Schule (Hellwig, 2012; Priemer & Hellwig, 2016) wollen wir für jedes der darin enthaltenen zehn Konzepte einen Test bereitstellen sowie die Zusammenhänge zwischen einzelnen Konzepten auf Basis unser empirisch erhobenen Daten analysieren (vgl. Tab. 1).

Grundsätzliche Existenz von Messunsicherheiten	**Ursachen der Messunsicherheit**	**Einfluss auf das Messwesen**	**Ziel der Messung**
	Unterscheidung zw. Messunsicherheit und Messabweichung		**Ergebnis der Messung**
Aussagekraft	**Verlässlichkeit der Messung und ihres Ergebnisses**	**Erfassung von Messunsicherheiten**	**Erfassung einer Unsicherheitskomponente bei direkter Messung**
	Vergleich von Messwerten		**Zusammensetzung der Messunsicherheit aus mehreren Komponenten**
	Regression		**Erweiterte Messunsicherheit**

Tabelle 1: Auszug aus dem Sachstrukturmodell (Hellwig 2012)

Forschungsfragen

In diesem Beitrag möchten Folgendes für die Konzepte „Verlässlichkeit der Messung und ihres Ergebnisses“ und „Vergleich von Messwerten“ untersuchen:

- Inwiefern lassen sich die Konzepte des Sachstrukturmodells in Form von Teilkompetenzen beschreiben, operationalisieren und messen?
- Welche Qualität haben die entwickelten Skalen auf Basis empirischer Daten einer Pilotierung?

Methode

Für die Entwicklung unseres Tests sind wir wie folgt vorgegangen: 1. Formulierung von Kompetenzen und deren Operationalisierung durch Testitems, 2. Expertenrating zur Validierung der Testitems und 3. empirischer Test der Items mit Schüler*innen.

Formulierung von Kompetenzen und Testitems
Für jedes der zehn Konzepte des Sachstrukturmodells haben wir auf Basis der Beschreibung der Inhalte sowie der Reduktionen aus (Hellwig, 2012; Priemer & Hellwig, 2016) Kompetenzen formuliert. Die Inhalte der zu den Konzepten gehörenden Subkonzepte aus dem oben genannten Modell wurden dabei ebenfalls berücksichtigt. Außerdem wurde darauf geachtet, mögliche Überschneidungen in den Inhalten der Konzepte auszuschließen. Exemplarisch sind in Tab. 2 die Subkonzepte für den Bereich „Verlässlichkeit der Messung“ und darunter die zugehörigen Kompetenzen aufgeführt.

Aussagekraft	
Verlässlichkeit der Messung und ihres Ergebnisses	Genauigkeit des Schätzwertes
	Grad des Vertrauens
	Rückschlüsse auf die Messung

Tabelle 2: Konzept „Verlässlichkeit der Messung“ aus Hellwig, 2012

Zu dem Konzept „Verlässlichkeit der Messung“ wurden die folgenden Kompetenzen formuliert: Die Schüler*innen können…
- stellen eine Unsicherheitsbilanz auf und analysieren diese hinsichtlich ihrer Vollständigkeit und der Stärke der Einflussfaktoren
- stellen eine Unsicherheitsbilanz auf und analysieren diese hinsichtlich ihrer Vollständigkeit und der Stärke der Einflussfaktoren
- stellen eine Unsicherheitsbilanz auf und analysieren diese hinsichtlich ihrer Vollständigkeit und der Stärke der Einflussfaktoren

Zu diesen Kompetenzen wurden dann Multiple Choice Testitems konstruiert. Diese enthielten sowohl Single Select als auch Multi Select Antwortmöglichkeiten. Bei den Multi Select Antworten war stets mindestens eine und niemals alle Antwortmöglichkeiten richtig. Die Kodierung erfolgte bei beiden Itemarten dichotom. Fehlende Antworten wurden als Missings angegeben. Insgesamt wurden für jedes der beiden hier diskutierten Konzepte 17 Items formuliert. Bei der Formulierung der Items wurde außerdem darauf geachtet, dass Inhalte verwendet werden, die Schüler*innen aus dem Rahmenlehrplan oder ihre Lebenswelt vertraut sind. Zusätzlich wurde den Items ein maximal einseitiger Informationstext mit den wichtigsten Regeln und Fachbegriffen des jeweiligen Konzeptes vorgestellt, um fehlende fachliche Inhalte bereitzustellen.

Expertenrating
Für das Expertenrating wurden drei Experten aus dem Bereich Messunsicherheiten 52 von insgesamt über 120 Testitems aus allen Konzepten vorgelegt. Diese wurden zufällig ausgewählt mit der Besonderheit, dass aus jedem Konzept mindestens ein Item im Expertenrating vorkommen sollte. Die Experten wurden dann gebeten, die Items den jeweiligen Konzepten auf Basis der formulierten Kompetenzen zuzuordnen. Zusätzlich wurde ein 11-tes Konzept für „nicht passende“ Items hinzugefügt. Darüber hinaus konnten die Experten auch einzelne Items direkt mit Kommentaren versehen.

Insgesamt haben die Experten 31 Items dem gleichen Konzept zugeordnet, 14 Items wurden von zwei Experten dem gleichen Konzept zugeordnet und 7 Items wurden in komplett verschiedene Konzepte einsortiert. Als Inter-Rater-Übereinstimmung ergab sich ein Fleiss´ Kappa von $\kappa = 0{,}67$. Für die sechs Items aus den beiden obigen Konzepten ergab sich ein Wert von $\kappa = 0{,}5$. Dies ergibt eine „moderate“ bis „substanzielle“ Übereinstimmung (nach Landis & Koch, 1977).

Pilotierung mit Schüler*innen

Zur Pilotierung wurden die Testitems insgesamt 143 Schüler*innen der Klassenstufe 8 - 12 aus sechs verschiedenen Berliner Schulen vorgelegt. Für die Bearbeitungsdauer gab es keine Beschränkung, insgesamt benötigte niemand länger als 90 Minuten. Um Ermüdungseffekte zu kompensieren, wurden in der Hälfte aller Testhefte die Reihenfolge der beiden Konzepte getauscht. Vor der Pilotierung wurden außerdem die Items gemäß den Anmerkungen der Experten überarbeitet und Testitems, die im Expertenrating in zwei oder mehr Kategorien fielen, überarbeitet.

Nach der Auswertung der erhobenen Daten mit R und Winsteps mit Hilfe der Item-Response-Theorie ergaben sich folgende EAP-Reliabilitäten: r = 0,54 für das Konzept „Verlässlichkeit der Messung“ und r = 0,8 für das Konzept „Vergleich von Messwerten“. Zur Überprüfung der Rasch-Konformität der Items wurden außerdem die Folgenden Parameter untersucht: MNSQ-Outfits, ICC-Plots, lokale statistische Unabhängigkeit, sub-group Invarianz, Homogenität der Items, Unsicherheiten der einzelnen Parameter. Im Ergebnis sind die Items weitestgehend Rasch-konform. Lediglich der MNSQ-Outfit eines Items lag außerhalb des von Linacre und Wright (1994) empfohlenen Intervalls von 0,7 - 1,3. Da alle Experten in diesem Item jedoch übereinstimmten und es u. E. ebenfalls inhaltskonform ist, haben wir das Item beibehalten. Die Wright-Maps der beiden Konzepte sind in Abb. 1 dargestellt.

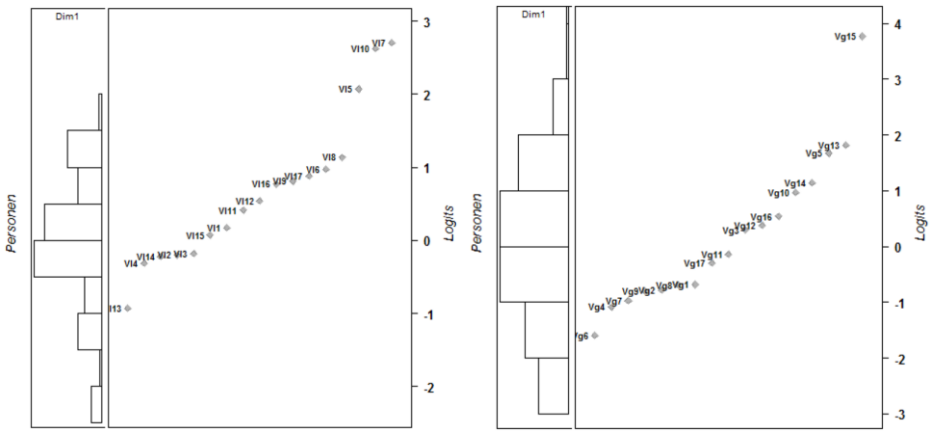

Abb. 1: Wright-Maps für die Konzepte „Verlässlichkeit der Messung“ (links) und „Vergleich von Messwerten“ (rechts)

Zusammenfassung und Ausblick

Insgesamt lässt sich sagen, dass die Messung der oben genannten Kompetenzen gut funktioniert und der Test in seiner Pilotversion akzeptabel ist. Zwar fehlen noch Items in bestimmten Schwierigkeiten, insgesamt wird jedoch die Fähigkeit der Schüler*innen gut getroffen und klare Unterschiede zwischen leichten und schwierigen Items sind erkennbar. Die Reliabilitäten sind hinreichend gut, wobei hier beachtet werden muss, dass aufgrund der im Sachstrukturmodell vorhandenen Subkonzepte Unterdimensionen auftreten können.

Aufbauend auf diesen Vorarbeiten (und anderen Pilotierungen weiterer Skalen des Modells) soll eine Hauptstudie mit ca. 750 Probanden folgen, um den Test mit allen zehn Dimensionen zu vervollständigen. Das dann erarbeitete Testinstrument soll perspektivisch die Möglichkeit bieten, Lernumgebungen und Instruktionen im Bereich der Messunsicherheiten in Bezug auf ihre Wirksamkeit zu untersuchen und Kompetenzzuwächse von Schüler*innen in den Konzepten des Modells zu erheben.

Literatur

BIPM, IEC, IFCC, ISO, IUPAC, IUPAP & OIML (1995/2008). Guide to the Expression of Uncertainty in Measurement (GUM). Geneva: International Organization for Standardization.

Buffler, A., Allie, S., Lubben, F. & Campbell, B. (2001). The development of first year physics students' ideas about measurement in terms of point and set paradigms. International Journal of Science Education, 23 (11), 1137-1156.

Day, J. & Bonn, D. (2011). Development of the Concise Data Processing Assessment. Physical review special topics - physics education research 7, 010114. Doi: 10.1103/PhysRevSTPER.7.010114

Deardorff, D. (2001). Introductory physics students' treatment of measurement uncertainty (Diss., North State University, Raleigh, NC). https://www.ncsu.edu/PER/Articles/DeardorffDissertation.pdf

Department for Education and Employment & Qualifications and Curriculum Authority (1999). Science The National Curriculum for England. Retrieved from http://dera.ioe.ac.uk/4402/1/cSci.pdf

Garratt, J., Horn, A. and Tomlinson, J. (2000). Misconceptions about error. University Chemistry Education, 4(2), 54-57.

Hellwig, J. (2012). Messunsicherheiten verstehen – Entwicklung eines normativen Sachstrukturmodells am Beispiel des Unterrichtsfaches Physik.
Dissertation: http://www-brs.ub.ruhr-unibochum.de/netahtml/HSS/Diss/HellwigJulia

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining change in psychotherapy research. Journal of Consulting and Clinical Psychology, 59, 12–19.

Klahr, D. & Dunbar, K. (1988). Dual Space Search During Scientific Reasoning. Cognitive Science, 12, 1-48. doi:10.1207/s15516709cog1201_1

KMK – Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2004). Bildungsstandards im Fach Physik für den Mittleren Schulabschluss [Science standards for middle school graduation for the school subject physics]. München: Wolters Kluwer.

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

Linacre, J. M. & Wright, B. D. (1994). Reasonable mean-square fit values, http://www.rasch.org/rmt/rmt83b.htm

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

Masnick, A. M., & Morris, B. J. (2008). Investigating the development of data evaluation: The role of data characteristics. Child Development, 79, 1032-1048. doi:10.1111/j.1467-8624.2008.01174.x.

Munier, V., Merle, H. & Brehelin, D. (2011). Teaching Scientific Measurement and Uncertainty in Elementary School. International Journal of Science Education, iFirst Article, 1-32, doi: 10.1080/09500693.2011.640360

NGSS Lead States (2013). Next generation science standards: For states, by states. Washington, DC: The National Academies Press.

Pedaste, Mäeots, Siiman, De Jong, Van Riesen, Kamp et al. (2015). Phases of inquiry-based learning: definitions and the inquiry cycle. Educational Research Review (14), 47-61, https://doi.org/10.1016/j.edurev.2015.02.003

Priemer, B. & Hellwig, J. (2016). Learning About Measurement Uncertainties in Secondary Education: A Model of the Subject Matter. International Journal of Science and Mathematics Education. doi:10.1007/s10763-016-9768-0.

Volkwyn, T. S. (2005), First year students' understanding of measurement in physics laboratory work. Dissertation at University of Cape Town.