Livia Murer[1]
Susanne Metzger[2]
Andreas Vorholzer[3]
Angela Bonetti[1]
Christoph Gut[1]

[1]Pädagogische Hochschule Zürich
[2]Fachhochschule Nordwestschweiz
[3]Justus-Liebig-Universität Gießen

Vergleich unterschiedlicher Methoden zur Einschätzung experimenteller Kompetenz im hands-on-Test

Experimentelle Kompetenz kann auf unterschiedliche Weise eingeschätzt werden, häufig mit paper-and-pencil-Tests, Computersimulationen oder hands-on-Tests (z. B. Schreiber et al., 2009; Baxter & Shavelson, 1994; Solano-Flores et al., 1999 oder Webb et al., 2000). Bei hands-on-Tests gibt es wiederum verschiedene *Methoden*, um die experimentelle Kompetenz zu schätzen, z. B. Beobachtungen, Schülerprotokolle oder Interviews. Studien zum Vergleich der Einschätzungen durch Beobachtungen und Schülerprotokolle kommen zu unterschiedlichen Ergebnissen bzgl. der Passung der Einschätzungen (z. B. Baxter & Shavelson, 1994; Shavelson et al., 1999; Hild et al., 2018 oder Schreiber et al., 2016). Shavelson et al. (1999) finden beispielsweise gute Übereinstimmungen zwischen Beobachtungen und Schülerprotokollen während bei Hild et al. (2018) die mit Beobachtungen und Schülerprotokollen geschätzten Kompetenzwerte nur gering miteinander korrelieren. Zum Vergleich von Schülerprotokollen und Interviews bzw. Interviews und Beobachtungen gibt es kaum Studien. Baxter et al. (1995) konnten zeigen, dass Lernenden, denen auf Basis einer Beobachtung ein hoher Kompetenzwert zugewiesen wurde, im Interview auch qualitativ bessere Erklärungen äußern und es somit anscheinend einen Zusammenhang zwischen diesen Methoden gibt.

Die vorgestellte Studie wurde im Rahmen des durch den Schweizerischen Nationalfonds geförderten Projekts ExKoNawi (*Ex*perimentelle *Ko*mpetenz in den *Na*tur*wi*ssenschaften) realisiert. Das im Projekt ExKoNawi entwickelte Modell zur Beschreibung experimenteller Kompetenz unterscheidet zwischen verschiedenen Problemtypen, z. B. «Naturwissenschaftliches Messen mit vorgegebenen Instrumenten», «Vergleichende Untersuchung von Objekten» und «Experimentelle Untersuchung der Beziehung zwischen Variablen» (Bonetti et al., 2016; Gut et al. 2014). Beim Problemtyp «Naturwissenschaftliches Messen mit vorgegebenen Instrumenten» (kurz «Messen») wird die experimentelle Kompetenz mit 15 Indikatoren eingeschätzt. Ziel dieser Studie ist es, bei den hands-on-Aufgaben des Problemtyps «Messen» verschiedene Methoden zur Einschätzung der experimentellen Kompetenz zu vergleichen. Dazu wird auf Basis von Schülerprotokollen (P), Beobachtungen mittels Videos (V) und Interviews (I) bewertet, inwiefern die 15 Indikatoren experimenteller Kompetenz von den Lernenden erfüllt werden. Die gemeinsame Betrachtung von P, V und I wird dabei als Benchmark betrachtet, der die genauste Schätzung experimenteller Kompetenz darstellt. Dabei gilt ein Indikator als erfüllt, sobald er von den Lernenden in einer der drei Methoden erfolgreich gezeigt wird. Mit dem Benchmark werden die Schätzung der einzelnen Methoden verglichen; die Forschungsfrage lautet:

Inwiefern gibt es systematische Abweichungen bei der Einschätzung der experimentellen Kompetenz beim Problemtyp «Messen» zwischen Schülerprotokollen (P), Schülerprotokollen und Beobachtungen (P&V), Schülerprotokollen und Interviews (P&I) und dem gesetzten Benchmark (P&V&I) bzgl. des geschätzten Kompetenzwertes?

Die Beobachtungen werden zusammen mit den Schülerprotokollen ausgewertet (P&V), da nicht alle Indikatoren im Video ersichtlich werden (z. B. Indikatoren zum Ergebnis werden im Video nicht ersichtlich). Auch die Interviews werden zusammen mit den Schülerprotokollen

ausgewertet (P&I), da die Protokolle bei den Interviews als Stimulus dienten und somit die Interviews nicht isoliert betrachtet werden können.

Stichprobe und Design

Die Stichprobe umfasst 27 Schülerinnen und Schüler (SuS) aus sechs verschiedenen Klassen (alle 8. Schuljahr; mittleres Niveau; 13-15 Jahre alt; 48% weiblich). Jede Klasse wurde vier Mal besucht. Bei jedem Besuch wurde eine andere, aber vergleichbare hands-on-Aufgabe des Problemtyps «Messen» von den SuS bearbeitet. Während dem Bearbeiten der Aufgabe haben die SuS ihre Vorgehensweise und Ergebnisse im Schülerprotokoll festgehalten (P) und wurden zusätzlich videografiert (V). Nach der Bearbeitung der Aufgabe wurden die SuS in Einzelinterviews dazu befragt, wie sie bei der Bearbeitung der Aufgabe vorgegangen sind und warum (I). Da jede Klasse insgesamt vier Mal besucht wurde, stehen insgesamt 108 Schülerprotokolle, 108 Videos und 108 Interviews für die Auswertung zur Verfügung.

Die Schülerprotokolle, Videos und Interviews wurden mit einem standardisierten Manual im Hinblick auf das Vorhandensein der 15 Indikatoren kodiert. Mind. 15 % der Daten (P, V bzw. I) wurden doppelt kodiert, mit zufriedenstellender Übereinstimmung ($\kappa \geq .61$; prozentuale Übereinstimmung ≥ 81 %).

Zur Beantwortung der Forschungsfrage wurden die Ergebnisse der einzelnen Methoden kombiniert (z. B. P&V, P&I oder P&V&I) und daraus Schätzwerte für die experimentelle Kompetenz der SuS ermittelt. Dabei wurde ein Indikator bei einer Kombination als erfüllt angesehen, wenn er – analog zum Vorgehen bei der Bildung des Benchmarks – in mind. einer der Methoden beobachtet werden kann. Wenn z. B. anhand des Schülerprotokolls nicht ersichtlich wird, ob Messwiederholungen gemacht wurden, im Interview aber gesagt wird, dass die Messungen wiederholt wurden, gilt der Indikator bei P&I als erfüllt. Anschließend wurden die mit den einzelnen Methoden geschätzten Kompetenzwerte auf Gruppen- und Individualebene mithilfe statistischer Tests verglichen.

Erste Ergebnisse

Beim Vergleich der Verteilung der geschätzten Kompetenzwerte bzw. der Mittelwerte der Schätzung durch P, P&V, P&I und P&V&I zeigt sich, dass die experimentelle Kompetenz durch P&I bzw. P&V&I tendenziell höher eingeschätzt wird als durch P bzw. P&V (Abb. 1 und Abb. 2). Zudem ist zu erkennen, dass die Methoden P&I und P&V&I zu einer sehr ähnlichen Schätzung des Kompetenzwertes kommen.

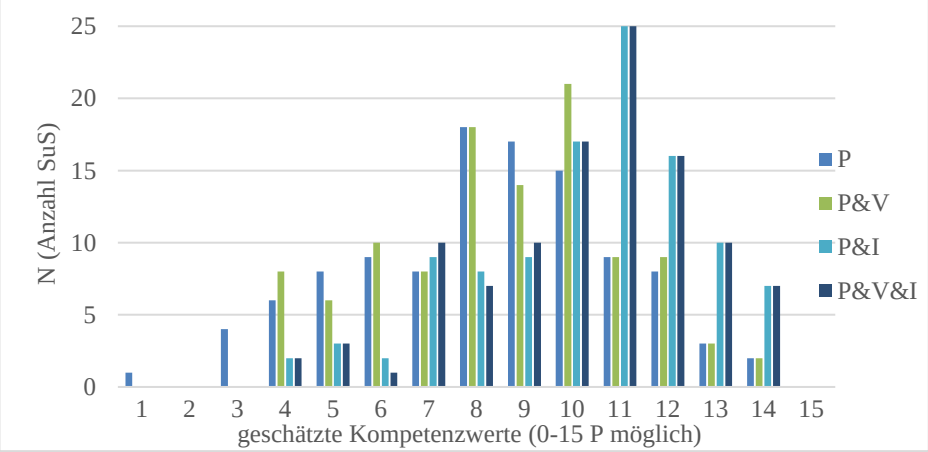
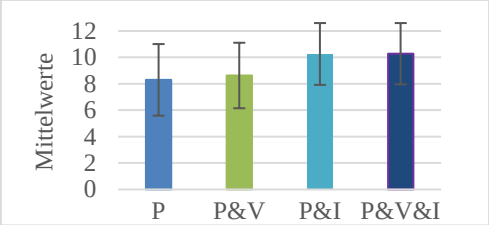


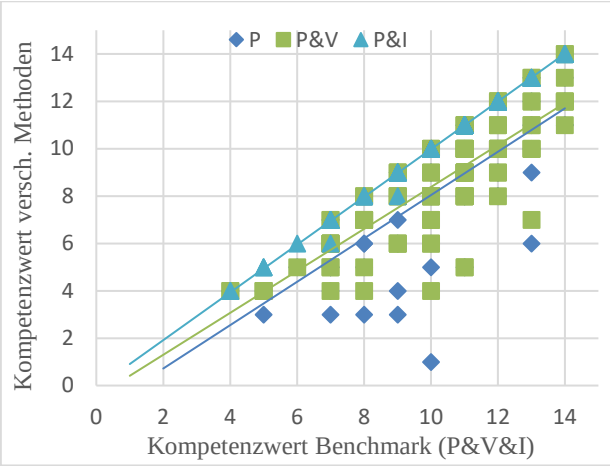
Abb. 1: Verteilung der geschätzten Kompetenzwerte durch P, P&V, P&I und dem Benchmark (P&V&I).



Mittelwertsunterschiede (t-Test abhängige Stichprobe)
- P vs. P&V&I: t = -12.14; p < .001; *N* = 108 (d = -.79)
- P&V vs. P&V&I: t = -12.09; p < .001; *N* = 108 (d = -.69)
- P&I vs. P&V&I: t = -1.42; n.s.

Abb. 2: Vergleich der Mittelwerte von P, P&V bzw. P&I und dem Benchmark (P&V&I). Fehlerbalken entsprechen +/- 1 SD.

Ein ähnliches Bild zeigt sich auf Individualebene, also wenn die Korrelationen zwischen den Kompetenzwerten für jede Person individuell betrachtet werden (Abb. 3). Die Korrelation zwischen den Methoden P&I und P&V&I ist sehr hoch. Somit scheinen die Kompetenzwerte durch P&I und P&V&I sehr ähnlich geschätzt zu werden. Weniger hoch ist die Korrelation zwischen den geschätzten Kompetenzwerten von P bzw. P&V und dem Benchmark (P&V&I).



Korrelationen (Kendall τ_b)
- P und P&V&I: r = .67; p < .001; *N* = 108
- P&V und P&V&I: r = .71; p < .001; *N* = 108
- P&I und P&V&I: r = .99; p < .001; *N* = 108

Abb. 3: Korrelationen und Regressionsgeraden zwischen P, P&V bzw. P&I und dem Benchmark (P&V&I).

Zusammenfassung und Ausblick

Die Ergebnisse deuten darauf hin, dass mithilfe der Interviews die experimentelle Kompetenz beim Problemtyp «Messen» genauer geschätzt werden kann als mit den Videos und/oder Schülerprotokollen alleine (geschätzte Kompetenzwerte durch P&I bzw. P&V&I sind höher als durch P bzw. P&V). Auch wenn SuS vermutlich über eine Teilkompetenz verfügen, werden die zugehörigen Indikatoren nicht immer im Video oder Protokoll sichtbar. Gerade im Protokoll scheint oft nur ein Teil des Vorgehens und der Ergebnisse notiert zu werden.

Die Schätzungen der experimentellen Kompetenz durch P&I kommen in der Regel sehr nahe an den gesetzten Benchmark (P&V&I) heran. Während das Interview eine wichtige Ergänzung zu Schülerprotokollen darstellt, scheint der zusätzliche Einsatz von Videos (V) bei diesem Messinstrument und diesem Problemtyp hingegen nicht zwingend notwendig zu sein. Bei der Auswertung konnten Unterschiede bzgl. des geschätzten Kompetenzwertes zwischen P bzw. P&V und P&I bzw. P&V&I gezeigt werden. Ergänzend soll nun analysiert werden, durch welche Indikatoren die Unterschiede hauptsächlich bedingt sind und wie das Auftreten dieser Unterschiede mit zentralen Personenvariablen zusammenhängt.

Literatur

Baxter, G. P. & Shavelson, R. J. (1994). Science performance assessments: Benchmarks and surrogates. *International Journal of Educational Research*, 21, 279-298.

Baxter, G. P., Elder, A. D. & Glaser, R. (1995). *Cognitive analysis of a science performance assessment* (CSE Technical Report Nr. 398). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, University of California.

Bonetti, A., Metzger S. & Gut C. (2016). Validierung des ExKoNawi-Modells (Experimentelle Kompetenzen in den Naturwissenschaften). *Implementation fachdidaktischer Innovation im Spiegel von Forschung und Praxis*, *37*, 332–335. Zürich.

Gut, C., Metzger, S., Hild, P., & Tardent, J. (2014). Problemtypenbasierte Modellierung und Messung experimenteller Kompetenzen. *PhyDid - Beiträge Zur DPG-Frühjahrstagung*.

Hild, P., Gut, C., & Brückmann, M. (2018). Validating performance assessments: measures that may help to evaluate students' expertise in 'doing science'. *Research in Science & Technological Education.*

Schreiber, N., Theyßen, H. & Schecker, H. (2009). Experimentelle Kompetenz messen?! *Physik und Didaktik in Schule und Hochschule*, 8 (3), 92-101.

Schreiber, N., H. Theyßen, and H. Schecker. (2016). Process-Oriented and Product-Oriented Assessment of Experimental Skills in Physics: A Comparison. In *Insights from Research in Science Teaching and Learning, Contributions from Science Education Research*, edited by N. Papadouris, A. Hadjigeorgiou, Angela, C. Constantinou, 29–43. Switzerland: Springer-Verlag.

Shavelson, R. J., Ruiz-Primo, M. A. & Wiley, E. W. (1999). Note on sources of sampling variability in science performance assessments. *Journal of Educational Measurement*, 36 (1), 61-71.

Solano-Flores, G., Jovanovic, J., Shavelson, R. J. & Bachman, M. (1999). On the development and evaluation of a shell for generating performance assessments. *International Journal of Science Education*, 21 (3), 293-315.

Webb, N. M., J. Schlackman, and B. Sugrue. (2000). The Dependability and Interchangeability of Assessment Methods in Science. *Applied Measurement in Education* 13 (3), 277–301.