Simon Schäfer[1] [1] Humboldt-Universität zu Berlin
Rüdiger Tiemann[1] [2] Universität Hamburg
Jenna Koenen[2]

Experimentieren in der Hochschule
Prüfung der Passung eines Modells

Erkenntnistheoretische Kompetenzen Der Prozess der Erkenntnisgewinnung ist eine zentrale Komponente des naturwissenschaftlichen Arbeitens im Speziellen und der Wissensgenese im Allgemeinen, der ein lebenslanges und selbstständiges Lernen ermöglicht (EuP & EuR, 30.12.2006). Daher sollen diese erkenntnistheoretische Kompetenzen als Teil einer naturwissenschaftlichen Grundbildung erworben werden, den das Vorhandensein dieser Kompetenzen ist notwendig für das Zurechtfinden in und die aktive Teilhabe an der uns umgebenen modernen Welt (KMK, 2004).

Nach der Wirkkette des Schulischen Lernens nach (Vogelsang 2014) müssen die bei Schülern*innen auszubildenden Kompetenzen auch bei den Lehrkräften hinreichend ausgeprägt sein. Um sicherzustellen, dass angehende Lehrkräfte die erforderliche Kompetenzausprägung erreichen, müssen auch in ihrer Ausbildungsphase entsprechende Lerngelegenheiten geboten und Instrumente zur Messung der Kompetenzausprägung eingesetzt werden (Biggs und Tang 2009). Der Einsatz solcher Testinstrumente würde es darüber hinaus ermöglichen, den Erfolg gezielter Interventionen zur Verbesserung des Kompetenzerwerbs im Bereich der Erkenntnisgewinnung in der Lehrerausbildung zu messen und zu bewerten.

Jedoch ist die Mehrzahl der die erkenntnistheoretischen Kompetenzen erhebenden Instrumente für den Einsatz im Sekundarschulbereich konzipiert[1]. Daher wurde im Rahmen einer Masterarbeit untersucht, inwieweit ein für Mittelstufenschüler*innen konzipiertes Instrument auch für die Kompetenzmessung bei Studierenden geeignet ist.

Testinstrument Das betrachtete Testinstrument ist von Andreas Nehring (2014) als multiple-choice-Test für Schüler*innen der Sekundarstufe I entwickelt und mit 612 Lernenden erprobt worden. Es basiert auf dem Modell zur Vernetzung der Erkenntnisgewinnung nach Nehring et al. (2013). Im Rahmen der Arbeit wurde auf das Experimentieren als eine der drei im theoretischen Modell beschriebenen Naturwissenschaftlichen Arbeitsweisen fokussiert.

Die im theoretischen Modell postulierte dreigliedrige Struktur des Wissenschaftlichen Denkens in Fragestellung und Hypothese (E1), Planung und Durchführung (E2) und Auswertung und Reflexion (E3) lässt sich in vielen anderen Modellen wiederfinden (vergleiche Emden und Sumfleth 2012). Für den Bereich Experimentieren wurden zehn Themenkomplexe mit jeweils drei Items konzipiert. Diese Items lassen sich jeweils einer der drei Phase des Wissenschaftlichen Denkens zuordnen.

Vorgehen Im Sommersemester 2016 wurde der Test 142 Studierenden aus Fach- und Lehramtsbachelor vorgelegt. Zur Reduzierung der Testzeit wurden den Studierenden jeweils zehn der dreißig zur Verfügung stehenden Aufgaben entsprechend dem balanced-incomplete- block-multi- matrix- design (Frey et al. 2009) vorgelegt. Der Testzeitpunk

[1] Eine Ausnahme bildet Stiller et al. (2015).

wurde dabei so gewählt, dass die Studierenden im Lehramtsbachelor noch keine expliziten Instruktionen zum Thema Erkenntnisgewinnung erhalten hatten.

Bei der Auswertung wurde zunächst den folgenden Fragen nachgegangen: 1) Inwieweit zeigen die Items und resultierenden Skalen adäquate statistische Kennwerte? 2) Liegt eine Passung zwischen psychometrisch bevorzugtem und theoretischen Modell vor?

Zur Beantwortung dieser Fragen wurden als Kriterien u. a. Kennwerte der Items wie Infit und Trennschärfe und die Reliabilität der Skalen überprüft. Items, die keine Modellpassung aufwiesen, wurden vor der Beurteilung der Skalengüte entfernt. Anhand der annähernd χ^2-verteilten Deviance (Walter und Rost 2011) der Modelle und unter Berücksichtigung von Informationskriterien wie AIC und BIC wurde der zweiten Frage entsprechend die Passung eines ein-, zwei- und dreidimensionalen Rasch-Modells untersucht. Hierzu wurde die Software Conquest (Adams et al. 2015) genutzt.

Des Weiteren wurde der Test auf Messinvarinaz geprüft, denn nur wenn der Test in den unterschiedlichen Subpopulationen dasselbe misst, sind die Ergebnisse der Erhebungen bei Schüler*innen und Studierenden vergleichbar. Hierzu wurde zunächst verglichen, ob mit Blick auf die empirischen Daten der beiden Erhebungen dieselbe psychometrische Struktur zu bevorzugen ist, da die Dimensionalität eines Tests bei vorliegender Messinvarianz erhalten bleiben muss (Levy und Svetina 2011).

Anschließend wurde mit Hilfe des R-Pakets eRm (Mair et al. 2016) ein Likelyhood-Quotienten-Test (LR-Test) durchgeführt, um die Hypothese der Gleichheit der Itemparameter in den beiden Subpopulationen zu überprüfen (Eid und Schmidt 2014). Mittels des nachfolgenden Wald-Testes wurde dann die Hypothese der Itemparameter-kongruenz für jedes Item einzeln getestet und die Ergebnisse graphisch in *Abb. 1* dargestellt. Hierbei wurde zunächst ein Rasch-Modell über alle 30 Items berechnet, da die Differenzmengen der Items mit ungenügenden Kennwerten beider Subpopulationen nicht leer waren.

Die im Wald-Test als problematisch identifizierten Items wurden aus dem Modell entfernt und das beschriebene Vorgehen erneut durchgeführt. Abschließend wurden die verbleibenden Items hinsichtlich ihrer Kennwerte geprüft, wobei diese nun dem in R berechneten Modell entnommen wurden. Zum Vergleich wurde das Verfahren noch einmal in umgekehrter Reihenfolge durchlaufen, sodass zuerst Items mit ungenügenden Kennwerten entfernt und dann auf gruppenabhängige Messvarianzen geprüft wurde.

Ergebnisse Die Beurteilung der Itemkennwerte nach den bei Bond und Fox (2015) und Tepner und Dollny (2014) angegebenen Schwellenwerten führte im zum Ausschluss von fünf Items[2], wobei nur die Unterschreitung der Trennschärfe von .3 relevant war. Es zeigt sich für die resultierende Skala für Studierenden eine EAP/PV-Reliabilität von .64. Nehring (2014) berichtet für die Skala bei Schüler*innen eine EAP/PV-Reliabilität von .66[3]. Die Items weisen also überwiegend zufriedenstellende Kennwerte auf und die resultierenden Skalen sind vergleichbar reliabel. Das theoretische dreigliedrige Modell konnte aus psychometrischer Perspektive in keiner der beiden Subpopulationen betätigt werden. In beiden Fällen ist die eindimensionale Struktur zu bevorzugen. Daher wurde im Weiteren mit den eindimensionalen Modellen gerechnet.

[2] Es handelt sich um die Items E1_1, E1_4, E1_5, E1_8 und E3_1
[3] Die Skalierung erfolgte unter Ausschluss der Items E2_1, E2_7, E2_8 und E3_4.

Die erste zu prüfende Voraussetzung für Messinvarianz, identische Dimensionalität, ist also erfüllt. Der LR-Test zeigt aber, dass für ein Rasch-Modell unter Verwendung aller Items keine Messinvarianz vorliegt (χ^2 (29) = 192.23, $p < .001$). Auch für Rasch-Modelle, die nur Items mit ausreichend guten Kennwerten einbeziehen, bescheinigt der LR-Test keine Messinvarianz. Der Wald-Test identifizierte 13 raschinhomogene Items (vgl. *Abb. 1*). Sowohl LR-Test als auch Wald-Test für ein um diese Items bereinigtes Rasch-Modell zeigen, dass dieser Schritt konsistent zu einem messinvarianten Aufgabenset führt.

Über die beiden Subgruppen verteilt wurden neun Items mit ungenügenden Itemkennwerten identifiziert. Fünf der neun Items wurden auch im Wald-Test für raschinhomogen befunden. In von den 13 raschinhomogenen bereinigten Rasch-Modell zeigten sich für das Item E1_1 eine signifikante Verbesserung der Itemkennwerte, sodass für ein finales Modell noch 14 Items zur Verfügung stehen. Beim Durchlaufen des umgekehrten Prozesses führte das Ausschließen der Items mit ungenügenden Kennwerten hingegen zu keiner signifikanten Änderung der Ergebnisse des LR- oder Wald-Tests. Es verblieben auf diese Weise also nur 13 Items für die Skalierung der Daten.

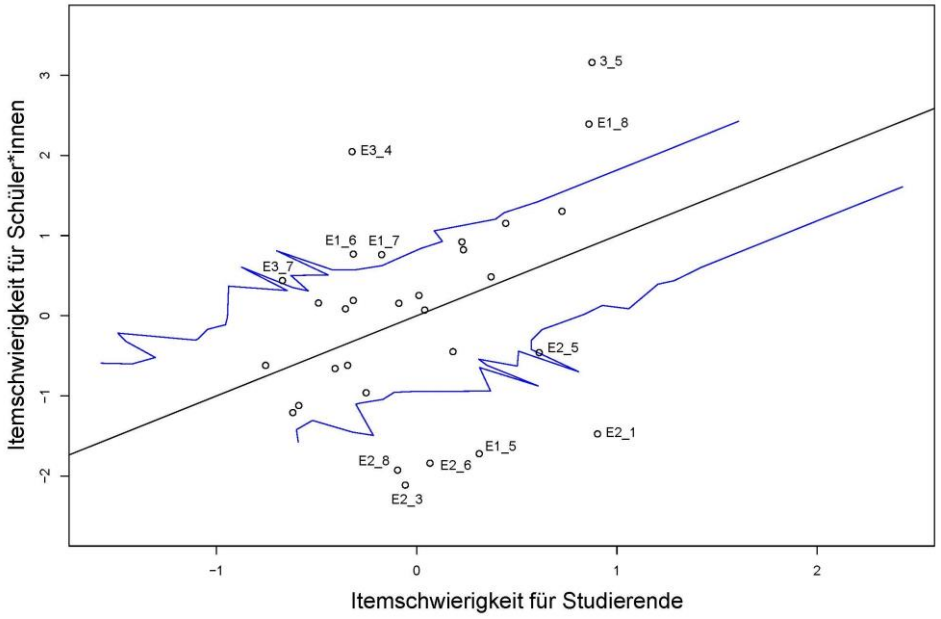

Abb. 1: Graphischer Modelltest. Beschriftete Items liegen außerhalb des Konfidenzbereichs und sind als nicht messinvariant einzustufen ($\alpha = .05$).

Diskussion und Ausblick Es wurde gezeigt, dass es einen Unterschied macht, in welcher Reihenfolge man die Items anhand ihrer Kennwerte oder des Ergebnisses des Wald-Tests entfernt. Eine Regel, die Aussagen trifft, welche Reihenfolge besser ist, lässt sich an diesem Einzelfall jedoch nicht ableiten. In weiteren Arbeitsschritten könnten die in beiden Studien erhobenen affektiven und kognitiven Kovariablen genutzt werden, um in einer konfirmatorischen Mehrgruppen-Faktorenanalyse die Zusammenhänge zum gemessenen Konstrukt zu untersuchen (Steinmetz et al. 2009). Abschließend sei die Frage aufgeworfen, inwieweit der Bedingung gleicher Dimensionalität für Messinvarianz im Falle von eindimensionalen Skalen Gewicht beigemessen werden kann, denn es gibt eine Vielzahl eindimensionaler Konstrukte. Sie ist sicherlich eine notwendige, aber nicht annähernd hinreichende Bedingung.

Literatur

Adams, R. J.; Wu, M. L.; Wilson, M. R. (2015): ACER ConQuest. Generalised Item Response Modelling Software. Version 4. Camberwell, Victoria: Australian Council for Educational Research.

Biggs, John Burville; Tang, Catherine (2009): Teaching for quality learning at university. What the student does. 3. ed., reprinted. Maidenhead: McGraw-Hill (McGraw-Hill education).

Bond, Trevor G.; Fox, Christine M. (2015): Applying the Rasch model. Fundamental measurement in the human sciences. Third edition. New York: Routledge Taylor & Francis Group. Online verfügbar unter http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&AN=1002030.

Eid, Michael; Schmidt, Katharina (2014): Testtheorie und Testkonstruktion. Göttingen: Hogrefe (Bachelorstudium Psychologie).

Emden, Markus; Sumfleth, Elke (2012): Prozessorientierte Leistungsbewertung - Zur Eignung einer Protokollmethode für die Bewertung von Experimentierprozessen. In: *MNU* 65 (2), S. 68–75.

Europäisches Parlament; Europäischer Rat (30.12.2006): Empfehlung des Europäischen Parlaments und des Rates vom 18. Dezember 2006 zu Schlüsselkompetenzen für lebensbegleitendes Lernen. 2006/962/EG. Online verfügbar unter https://www.bmb.gv.at/schulen/ejid/eu_amtsblatt_schlkomp_15538.pdf?5h6xww, zuletzt geprüft am 13.10.2016.

Frey, Andreas; Hartig, Johannes; Rupp, André A. (2009): An NCME Instructional Module on Booklet Designs in Large-Scale Assessments of Student Achievement. Theory and Practice. In: *Educational Measurement: Issues and Practice* 28 (3), S. 39–53. DOI: 10.1111/j.1745-3992.2009.00154.x.

Kultusministerkonferenz (KMK) (2004): Bildungsstandards im Fach Chemie für den Mittleren Schulabschluss. Beschluss vom 16.12.2004. Hg. v. Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. München, Neuwied. Online verfügbar unter https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2004/2004_12_16-Bildungsstandards-Chemie.pdf, zuletzt aktualisiert am 24.06.2008, zuletzt geprüft am 10.09.2016.

Levy, Roy; Svetina, Dubravka (2011): A generalized dimensionality discrepancy measure for dimensionality assessment in multidimensional item response theory. In: *The British journal of mathematical and statistical psychology* 64 (Pt 2), S. 208–232. DOI: 10.1348/000711010X500483.

Mair, P.; Hatzinger, R.; Maier, M. J. (2016): eRm. Extended Rasch Modelin. Version 0.15-7. Online verfügbar unter http://erm.r-forge.r-project.org/.

Nehring, Andreas (2014): Wissenschaftliche Denk- und Arbeitsweisen im Fach Chemie. Eine kompetenzorientierte Modell- und Testentwicklung für den Bereich der Erkenntnisgewinnung. Berlin: Logos Verl. (Studien zum Physik- und Chemielernen, 177).

Nehring, Andreas; Nowak, Kathrin Helena; Tiemann, Rüdiger; Upmeier zu Belzen, Annette (2013): Assessing students' abilities in processes of scientific inquiry in biology using a paper-and-pencil test. In: *Journal of Biological Education* 47 (3), S. 182–188. DOI: 10.1080/00219266.2013.822747.

Steinmetz, Holger; Schmidt, Peter; Tina-Booh, Andrea; Wieczorek, Siegrid; Schwartz, Shalom H. (2009): Testing measurement invariance using multigroup CFA. Differences between educational groups in human values measurement. In: *Qual Quant* 43 (4), S. 599–616. DOI: 10.1007/s11135-007-9143-x.

Tepner, Oliver; Dollny, Sabrina (2014): Entwicklung eines Testverfahrens zur Analyse fachdidaktischen Wissens. In: Dirk Krüger, Ilka Parchmann und Horst Schecker (Hg.): Methoden in der naturwissenschaftsdidaktischen Forschung. Berlin: Springer Spektrum, S. 311–323.

Vogelsang, Christoph (2014): Validierung eines Instruments zur Erfassung der professionellen Handlungskompetenz von (angehenden) Physiklehrkräften. Zusammenhangsanalysen zwischen Lehrerkompetenz und Lehrerperformanz. Zugl.: Paderborn, Univ., Diss., 2014. Berlin: Logos Berlin (Studien zum Physik- und Chemielernen, 174).

Walter, Oliver; Rost, Jürgen (2011): Psychometrische Grundlagen von Large Scale Assessments. In: Lutz Hornke (Hg.): Methoden der psychologischen Diagnostik. Göttingen [u.a.]: Verl. für Psychologie, Hogrefe (Psychologische Diagnostik, Bd. 2), S. 87–149.