

Laura Behrends
 Laura Muth
 Roger Erb

Goethe-Universität Frankfurt

Validierung eines Testinstruments zur Auswertekompetenz

Einleitung und Motivation

Um die Experimentierkompetenz von Schülerinnen und Schülern zu erfassen, werden reale Experimentiersituationen, Computersimulationen oder Paper-Pencil-Tests eingesetzt. Paper-Pencil-Tests sind zwar am wenigsten aufwändig, werden aber auch am problematischsten eingeschätzt, wenn es darum geht, Handlungskompetenzen zu erfassen (Schreiber, 2012). In einer Studie von Muth (2016), in der die Auswertephase von Experimenten im Physikunterricht im Fokus steht, sollen aus testökonomischen Gründen die Auswertekompetenzen der Schülerinnen und Schüler mittels eines Multiple-Choice-Tests erhoben werden. Die Auswertung von Experimenten erfolgt meist schriftlich. Daher stellt sich die Frage, ob ein Paper-Pencil-Test in der Lage ist, die relevanten Auswertekompetenzen von Schülerinnen und Schülern zu erfassen.

Ziel dieser Arbeit war es, die Validität des von Muth (2016) eingesetzten Fragebogens zur Auswertekompetenz zu ermitteln. Durch die Kriteriumsvalidität wird eine Aussage über die Übereinstimmung zwischen Testinstrument und Außenkriterium getroffen. Als Außenkriterium wurden Schülerprotokolle herangezogen. Diese wurden mittels quantitativer Inhaltsanalysen hinsichtlich verschiedener Kategorien bewertet (ob bspw. eine Skizze angefertigt wurde) und anschließend mit den Daten aus den Fragebögen verglichen. Anhand der Übereinstimmung dieser Daten wurde ein Korrelationskoeffizient berechnet, der eine Aussage über die Kriteriumsvalidität der einzelnen Kategorien und des gesamten Fragebogens trifft. Aus den Fragebogenergebnissen kann außerdem geschlossen werden, für wie relevant die Schülerinnen und Schüler die Umsetzung der einzelnen Kategorien in der Auswertung einschätzen. Dies ermöglicht einen Vergleich zwischen Relevanz aus Sicht der Schülerinnen und Schüler und Validität. Es liegt dabei nahe, dass besonders Kategorien, die eine hohe Relevanz besitzen, sich in den Protokollen wiederfinden.

Theoretischer Hintergrund

Die Studie von Muth (2016) untersucht, inwiefern die Auswertephase von Experimenten im Physikunterricht Einfluss auf den Zuwachs von Fachwissen und Auswertekompetenz hat. Diese wurde unter anderem über Fertigkeiten von vier Auswertemethoden „Skizze“, „Wertetabelle“, „Graph“ und „Schriftliches Auswerten“ operationalisiert. Genauere Ausführungen finden sich bei Muth und Erb (2017, in Vorbereitung). Die Auswertekompetenz wurde dabei mit einem schriftlichen Fragebogen erfasst, der unter anderem auf Basis des MeK-LSA Projekts (Dickmann et al., 2013) eigens entwickelt wurde. Einige weitere Items verlangten von den Schülerinnen und Schülern eine Einschätzung, ob bestimmte Kategorien (bspw. das Anfertigen einer Skizze) wichtig für eine Auswertung sind. Hieraus wurde die Relevanz der Kategorie berechnet.

In der vorliegenden Arbeit galt es zu überprüfen, ob dieser Fragebogen die tatsächlichen Auswertekompetenzen der Schülerinnen und Schüler valide erfasst. Als eines der drei Hauptgütekriterien psychologischer Tests steht somit die Validität im Mittelpunkt der vorliegenden Studie. Diese gibt an, ob ein Test das Konstrukt, das gemessen werden soll (Zielkonstrukt), auch tatsächlich erfasst. Unterschieden wird dabei in Inhalts-, Konstrukt- und Kriteriumsvalidität. Für diese Studie wurde die Kriteriumsvalidität untersucht, da durch die Erhebung eines Außenkriteriums auf das Auswerteverhalten der Schülerinnen und Schüler außerhalb der Testsituation geschlossen werden kann (Moosbrugger & Kelava,

2012). Ein Vergleich von Testinstrument und Außenkriterium ermöglicht es einen Korrelationskoeffizienten zu ermitteln und so eine Aussage über die Validität zu treffen. Dabei kann nach Fisseni (1990) ab einem Wert von 0,4 von einer mittleren und ab 0,6 von einer hohen Validität gesprochen werden. Zusätzlich zur Validität wurde mit Hilfe einiger Items die Relevanz der Kategorien erhoben. Diese gibt an, für wie wichtig Schülerinnen und Schüler bestimmte Aspekte bei der Auswertung einschätzen.

Design der Studie

Im Rahmen der vorliegenden Studie wurde folgenden Forschungsfragen nachgegangen:

F1: Lässt sich die Auswertekompetenz von Schülerinnen und Schüler valide mit einem Paper-Pencil-Verfahren erfassen?

F2: Weisen Kategorien mit einer als hoch eingeschätzten Relevanz eine hohe Validität auf?

Um diese Fragen zu beantworten, wurden Daten in mehreren Klassen innerhalb einer Doppelstunde erhoben. Zunächst wurde der schriftliche Fragebogen zur Auswertekompetenz von den Schülerinnen und Schüler ausgefüllt. Im Anschluss daran erfolgte ein Schülerexperiment in Kleingruppen, welches nachfolgend in Einzelarbeit von diesen - mithilfe einer Protokollvorlage - ausgewertet wurde.

Methode

Validitätsprüfung

Überprüft wurde die Kriteriumsvalidität des Fragebogens in der hier vorliegenden Arbeit mithilfe der quantitativen Inhaltsanalyse. Dabei wurde sich auf Mayring (2015) berufen, um die generierten Daten anhand eines Codebuches zu kategorisieren und mittels Korrelationskoeffizienten quantitativ auszuwerten. Als Außenkriterium wurden die erstellten Schülerprotokolle herangezogen. In einer Pilotierungsphase wurden dann zu denen im Fragebogen definierten Auswertemethoden „Skizze“, „Wertetabelle“, „Graph“ und schriftliche Auswertung („Schriftliches“) Kategorien gebildet, die sich in beiden Datensätzen wiederfanden (Mayring, 2015). Bei den daraus resultierenden Kategorien handelt es sich um Aspekte der jeweiligen Auswertemethode, die für die Schülerinnen und Schüler zentral sein könnten, wie beispielsweise das Zeichnen einer Skizze mit Bleistift.

Skizze	Wertetabelle	Graph	Schriftliches
Beschriftung	Messwerte	Koordinatensystem	Beobachtungen
Materialien	Physikalische Größen	Farben	Vermutungen
Bleistift	Lineal	Skalenbeschriftung	Verallgemeinerungen
Anordnung	Bleistift	Einheit	
Beobachtungen	Einheit	Lineal	
Farben		Achsenbeschriftung	
		Messwerte	

Tab. 1: Kategorien der Inhaltsanalyse

Um anhand der Datensätze eine quantitative Aussage über die Validität des Tests treffen zu können, wurde folgendes Bewertungssystem verwendet:

- Trifft eine Kategorie sowohl im Fragebogen, als auch im Protokoll zu, wurde dieser Übereinstimmung ein Übereinstimmungspunkt (ÜP) zugeschrieben.
- Trifft die Kategorie in keinem der Fälle zu, wurde dies ebenfalls als Übereinstimmung gewertet, wodurch ein ÜP zugeschrieben wurde.
- Wird die Kategorie nur bei einem der beiden Fälle genannt, herrscht keine Übereinstimmung, und es wird kein ÜP vergeben.

Anhand dieser ÜP kann ein Korrelationskoeffizient zwischen Aussagen des Fragebogens und der Schülerprotokolle errechnet werden.

Relevanzprüfung

Aus den Angaben einiger bestimmter Items konnte zudem die von den Schülerinnen und Schüler eingeschätzte Relevanz der Kategorien bestimmt werden. Die Schülerinnen und Schüler ordneten hierfür die vorgegebenen Aspekte nach ihrer Bedeutung in einer Reihenfolge an.

Ergebnisse

Die Studie wurde mit zwei Gymnasialklassen aus dem Rhein-Main-Gebiet durchgeführt. Bei den Klassen handelte es sich um eine siebte und eine neunte Jahrgangsstufe, zusammengesetzt aus insgesamt 11 Schülerinnen und 28 Schülern. Bei der Untersuchung der Auswertekategorien, ergaben sich folgende Koeffizienten:

Skizze	0,58	Wertetabelle	0,78	Graph	0,83	Schriftliches	0,63
Beschriftung	0,38	Messwerte	1	Koordinatensystem	0,83	Beobachtungen	0,9
Materialien	0,42	Physikalische Größen	1	Farben	0,95	Vermutungen	0,49
Bleistift	0,63	Lineal	0,56	Skalenbeschriftung	0,81	Verallgemeinerungen	0,49
Anordnung	0,67	Bleistift	0,56	Einheit	0,83		
Beobachtungen	0,61	Einheit	0,77	Lineal	0,6		
Farben	0,79			Achsenbeschriftung	0,83		
				Messwerte	0,93		

Tab. 2: Korrelationskoeffizienten der Auswertekategorien

Der geringste Korrelationskoeffizient ergab sich innerhalb der Auswertemethode „Skizze“, er liegt dabei knapp unter dem Schwellenwert für eine hohe Validität. Innerhalb der Kategorien besitzt die „Beschriftung“ einer Skizze den niedrigsten Wert. Mit Ausnahme dieser Kategorie überschreiten alle anderen den Schwellenwert von 0,4 und sind damit mindestens einer mittleren Validität zuzuordnen.

Anhand der Korrelationskoeffizienten der einzelnen Kategorien, lässt sich ebenfalls ein Korrelationskoeffizient für den gesamten Fragebogen errechnen. Dieser beläuft sich auf einen Wert von 0,72. Die Validität kann somit als hoch bewertet werden.

Vergleich von Relevanz und Validität

Bis auf eine Ausnahme lässt zeigt sich, dass Kategorien mit einer hohen Relevanz auch eine hohe Validität aufweisen. Lediglich die „Beschriftung“ einer Skizze wird zwar von den Schülerinnen und Schülern als sehr relevant eingeschätzt, jedoch nicht im Protokoll angewendet, diese Diskrepanz sorgt für eine geringe Validität. Auch Kategorien mit einer besonders niedrigen Relevanz zeigen erwartungskonform eine hohe Validität.

Diskussion und Ausblick

Die Ergebnisse zeigen, dass der eingesetzte Fragebogen eine hohe Validität besitzt, zumindest im Rahmen der hier angewandten Prüfung.

Aus dem Vergleich von Relevanz und Validität lässt sich ablesen, dass in den meisten Fällen die Schülerinnen und Schüler ihre Überzeugung hinsichtlich der verschiedenen Aspekte der Auswertemethoden auch im Protokoll umsetzen. Dennoch scheint gerade das Beschriften einer Skizze, ein aus Sicht der Lernenden wichtiger Aspekt zu sein, auf den sie in ihrer eigenen Auswertung häufig verzichteten.

In Anlehnung an diese Arbeit könnte es für nachfolgende Studien interessant sein, die Relevanz der einzelnen Kategorien von Lehrkräften einschätzen zu lassen und mit den hier erhobenen Daten zur Relevanz aus Sicht der Schülerinnen und Schüler zu vergleichen sowie die quantitative Inhaltsanalyse als Validierungsmethode mit einer größeren Stichprobe weiter zu untersuchen.

Literatur

- Dickmann, Martin, Eickhorst, Bodo, Theyßen, Heike, Neumann, Knut, Schecker, Horst. & Schreiber Nico (2013). Measuring experimental skills in large-scale assessments: developing a simulation-based test instrument. In C. P. Constantinou, N. Papadouris & A. Hadjigeorgiou (Eds.), *Science Education Research For Evidence-based Teaching and Coherence in Learning*. Proceedings of the ESERA 2013 Conference. Fisseni, Hermann-Josef (1990): Lehrbuch der psychologischen Diagnostik. Göttingen: Verl. für Psychologie Hogrefe.
- Mayring, Philipp (2015): Qualitative Inhaltsanalyse. Grundlagen und Techniken. 12., überarb. Aufl. Weinheim: Beltz (Beltz Pädagogik). Online verfügbar unter http://content-select.com/index.php?id=bib_view&ean=9783407293930.
- Moosbrugger, Helfried; Kelava, Augustin (2012): Testtheorie und Fragebogenkonstruktion. 2., aktualisierte und überarbeitete Auflage. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg (Springer-Lehrbuch). Online verfügbar unter <http://dx.doi.org/10.1007/978-3-642-20072-4>.
- Muth, Laura; Erb, Roger (2016): Einfluss der Auswertephase von Experimenten im Physikunterricht auf den Fachwissenszuwachs von Schülerinnen und Schülern. In: Christian Maurer (Hg.): Implementation fachdidaktischer Innovation im Spiegel von Forschung und Praxis. Jahrestagung. Zürich. Gesellschaft für Didaktik der Chemie und Physik (GDChP), S. 296–299.
- Schreiber, Nico (2012): Diagnostik experimenteller Kompetenz. Validierung technologiegestützter Testverfahren im Rahmen eines Kompetenzstrukturmodells. Zugl.: Duisburg, Essen, Univ., Diss., 2012. Berlin: Logos-Verl. (Studien zum Physik- und Chemielernen, 139).