

2 Modellierung der Testletstruktur bei vignettenbasierten Testverfahren mit geschlossenem Antwortformat

Juliane Rutsch, Markus Vogel, Markus Rehm & Tobias Dörfler

Zusammenfassung

Aktuelle Ansätze der empirischen Bildungsforschung verwenden Unterrichtsvignetten als Testaufgaben zur Erfassung professioneller Kompetenzen bei Lehrkräften. Für die psychometrische Modellierung vignettenbasierter Testverfahren ergeben sich anspruchsvolle, methodische Implikationen, da diese eine sogenannte Testletstruktur aufweisen. Dieser Beitrag untersucht zwei methodische Ansätze zur Berücksichtigung einer vorhandenen Testletstruktur für einen Vignettentest mit geschlossenem Antwortformat. Folgerungen für die Verwendung dieser beiden Ansätze zur Auswertung von Vignettentests mit geschlossenem Antwortformat werden diskutiert.

2.1 Theoretischer Hintergrund

Zur kontextualisierten und situierten Erfassung professioneller Kompetenzen von (angehenden) Lehrkräften, haben sich sog. vignettenbasierte Testverfahren bewährt (Blömeke, König, Suhl, Hoth & Döhrmann, 2015; Brovelli, Bölsterli, Rehm & Wilhelm, 2014; Tepner & Dollny, 2014). Diese Testverfahren zeichnen sich dadurch aus, dass sie Unterrichtsvignetten als Testaufgaben verwenden. Unterrichtsvignetten sind kurze, authentische Darstellungen von Situationen aus dem schulischen (Fach-)Unterricht, die als Text oder als Video präsentiert werden (Rehm &

Bölsterli, 2014). Durch die Verwendung von Unterrichtsvignetten, soll einer zentralen, methodischen Herausforderung an die empirische Erfassung professioneller (Lehrer-)Kompetenzen begegnet werden, nämlich einen authentischen Bezugsrahmen für die Kompetenzerfassung zu berücksichtigen (Shavelson, 2013).

Überschrift						
Kurze Beschreibung einer Unterrichtssituation in Textform oder als Videosequenz.						
Fachdidaktisch relevante Fragestellung						
	Trifft gar nicht zu					Trifft voll zu
Die Lehrkraft sollte...	[]	[]	[]	[]	[]	[]
Die Lehrkraft sollte...	[]	[]	[]	[]	[]	[]
Die Lehrkraft sollte...	[]	[]	[]	[]	[]	[]
Die Lehrkraft sollte...	[]	[]	[]	[]	[]	[]

Abbildung 2.1: Schematische Darstellung einer Unterrichtsvignette.

Verschiedene Studien, die Unterrichtsvignetten als Testaufgaben zur Erfassung professioneller Lehrkompetenzen herangezogen haben, verwenden offene Antwortformate (Baer & Buholzer, 2005; Brovelli, Bölsterli, Rehm & Wilhelm, 2013; Pissarek & Schilcher, 2015). Um eine ökonomische Datenerfassung und -auswertung für vignettenbasierte Testverfahren zu realisieren, gibt es zudem erste Bemühungen vignettenbasierte Testverfahren mit ausschließlich geschlossenem Antwortformat zu entwickeln (Tepner & Dollny, 2014). Das Forschungsprojekt EKoL greift diesen methodischen Ansatz auf: Es wurden domänenübergreifend vignettenbasierte Testverfahren mit geschlossenem Antwortformat entwickelt, um zu

überprüfen, ob sich dieser Ansatz als geeignet erweist, um professionelle Kompetenzen bei Lehrkräften zu erfassen.

Für einen Vignettentest mit geschlossenem Antwortformat werden für jede Unterrichtssituation mehrere Items präsentiert, die von den Probanden hinsichtlich einer konkreten Fragestellung eingeschätzt werden sollen (siehe Abbildung 2.1).

Die Bewertungen der Probanden werden anschließend mit den Bewertungen einer Expertengruppe verglichen (auch „Expertennorm“, „aggregierter Experte“, „Masterrating“, vgl. Oser & Forster-Heinzer, 2015; Oser, Heinzer & Salzmann, 2010). Detaillierte Ausführungen zur Erstellung einer Expertennorm für vignettenbasierte Testverfahren finden sich in Kapitel 7 dieses Bandes.

Für die Punktvergabe für Unterrichtsvignetten mit geschlossenem Antwortformat können anhand einer Expertennorm zwei Möglichkeiten in Betracht gezogen werden: Entweder eine Punktvergabe auf Itemebene (Meschede, Steffensky, Wolters & Möller, 2015) oder eine Punktvergabe anhand von Itemrelationen (Artelt, Beinicke, Schlagmüller & Schneider, 2009; Tepner & Dollny, 2014).

Für die Punktvergabe auf Itemebene wird für jedes Item geprüft, ob der Proband dieses genauso bewertet hat wie die Expertengruppe (siehe Abbildung 2.2). Hat der Proband ein Item beispielsweise auf einer sechsstufigen Skala mit „2“ bewertet und die Experten ebenso, dann erhält der Proband für diese Bewertung einen Punkt. Hat der Proband nicht die „2“ gekreuzt und unterscheidet sich damit von der Bewertung durch die Expertennorm, erhält er dafür null Punkte. Alternativ sind auch Punktvergaberegeln denkbar, die bei Übereinstimmung des Wertes des Probanden mit der Expertennorm einen Punkt und Teilpunkte (z. B. 0.5 Punkte) vergeben, wenn der Proband nahe der Expertennorm kreuzt, diese aber nicht erreicht.

Bei einer Punktvergabe anhand von Itemrelationen (auch: „Paarvergleich“) wird das Verhältnis, in dem zwei Items zueinander eingeschätzt wurden, bewertet. Wenn der Proband z. B. die Handlungsalternative in Item a) als zutreffender bewertet, als die Handlungsalternative in Item b) und laut Expertenurteil ist die Relation beider Items zueinander ebenso einzuschätzen, dann erhält der Proband für diese Itemrelation einen Punkt. Dabei ist unerheblich, an welcher absoluten Position auf der Skala der Proband seine Kreuze gesetzt hat (siehe Abbildung 2.3). Bei

der Punktvergabe anhand von Itemrelationen wird also bewertet, ob die Probanden verschiedene konkurrierende Handlungsalternativen gegeneinander so abwägen, wie es Experten tun.

Trifft gar nicht zu					Trifft voll zu
[]	<input checked="" type="checkbox"/>	[]	[]	[]	[]
X = Expertennorm X = Studierendenantwort					

Trifft gar nicht zu					Trifft voll zu
[]	<input checked="" type="checkbox"/> →	<input checked="" type="checkbox"/>	[]	[]	[]
X = Expertennorm X = Studierendenantwort					

Trifft gar nicht zu					Trifft voll zu
[]	<input checked="" type="checkbox"/> →	[] →	<input checked="" type="checkbox"/>	[]	[]
X = Expertennorm X = Studierendenantwort					

Abbildung 2.2: Darstellung der Punktvergabe auf Itemebene. Oben ist dargestellt, dass der Proband das Item genauso bewertet wie die Experten; hierfür erhält der Proband einen Punkt. In der Mitte der Abbildung ist dargestellt, dass der Proband das Item mit einer Abweichung von $|x| = 1$ ankreuzt und somit hierfür 0.5 Punkte

erhält. Unten ist die Situation illustriert, in der ein Proband $|x| > 1$ ankreuzt und hierfür keinen Punkt bekommt.

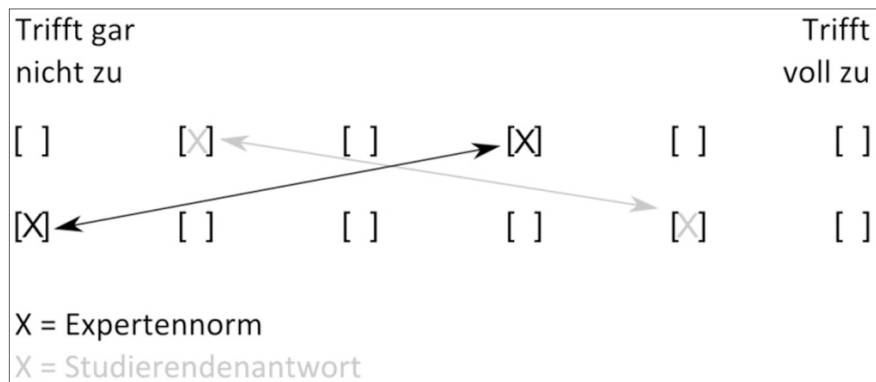
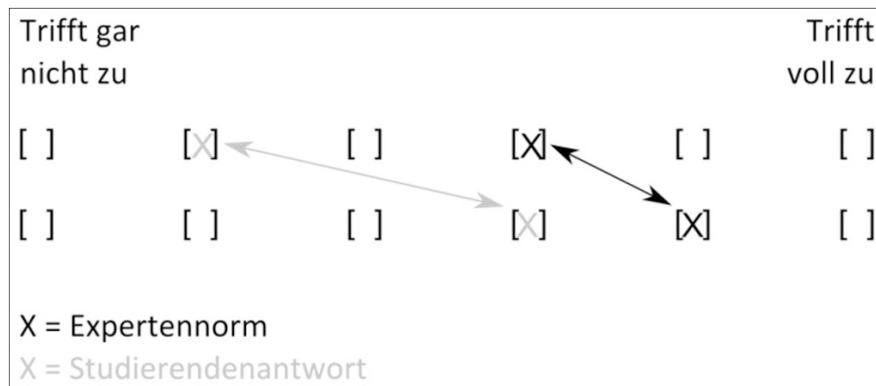


Abbildung 2.3: Darstellung der Punktvergabe anhand von Itemrelationen. Oben ist dargestellt, dass der Proband zwei Items in ihrer Relation zueinander ebenso bewertet wie die Expertennorm; dabei ist die absolute Position der Antwort auf der Skala nicht bedeutsam. Der Proband erhält hierfür einen Punkt. Im unteren Bereich der Abbildung ist dargestellt, dass der Proband eine andere, hier entgegengesetzte, Itemrelation gekreuzt hat als der aggregierte Experte; hierfür erhält der Proband 0 Punkte.

Wie sich diese beiden unterschiedlichen Punktvergabesysteme auf die Validität und die Reliabilität von vignettenbasierten Testverfahren auswirken, ist bislang noch nicht untersucht worden und stellt daher ein Forschungsdesiderat für weiterführende Analysen dar (siehe auch Diskussion).

Unabhängig von Fragen zur angemessenen Punktvergabe für Unterrichtsvignetten ergeben sich bei der statistischen Modellierung von vignettenbasierten Testverfahren besondere methodische Erfordernisse, die zu beachten sind. Diese sind darauf zurückzuführen, dass ein Vignettentest in der Regel eine sog. *Testletstruktur* aufweist. Von einer Testletstruktur wird gesprochen, wenn mehrere Items innerhalb eines Testinstruments gruppiert vorliegen (Bühner, 2011; Wainer & Kiely, 1987). Für das Beispiel „Vignettentest“ bedeutet das, dass die Items, die zu den jeweiligen Vignetten präsentiert werden, als gruppiert angesehen werden müssen; diese Items beziehen sich nämlich auf denselben Vignettenstamm und weisen daher eine inhaltliche und statistische Abhängigkeit voneinander auf.

Für die Modellierung von Daten in Item-Response-Modellen wird vorausgesetzt, dass die eingesetzten Testaufgaben (hier: Vignetten), abgesehen vom zu erfassenden latenten Merkmal, unabhängig voneinander sind (für einen Überblick: Strobel, 2012). Die verwendeten Testaufgaben dürfen daher nach Herabspaltung des zu erfassenden latenten Merkmals nicht miteinander oder mit weiteren externen Merkmalen korrelieren (Embretson & Reise, 2000). Das bedeutet, dass für die Modellierung von Item-Response-Modellen lokale, stochastische Unabhängigkeit vorausgesetzt wird (Massof, 2011). Hierunter wird verstanden, dass sich die Lösungswahrscheinlichkeit für eine Testaufgabe nicht dadurch verändern darf, dass zuvor eine andere Testaufgabe gelöst wurde (siehe auch: Strobel, 2012). Im vorliegenden Beispiel darf sich also die Lösungswahrscheinlichkeit für eine Vignette nicht dadurch erhöhen, dass zuvor eine ähnliche Vignette bearbeitet wurde. Bei der Testkonstruktion ist daher u. a. darauf zu achten, dass die Testaufgaben nicht inhaltlich aufeinander aufbauen.

Wird eine vorliegende Testletstruktur bei der statistischen Modellierung nicht beachtet, kann dies u. a. zu einer Überschätzung der Reliabilität bzw. Verzerrung der Testinformationsfunktion für das Testverfahren führen (Wainer & Wang,

2000). Zur psychometrischen Berücksichtigung einer vorhandenen Testletstruktur in einem Testinstrument haben sich zwei unterschiedliche Ansätze als geeignet herausgestellt (Wilson & Adams, 1995):

Score-basierte Ansätze addieren die Items innerhalb einer inhaltlichen Einheit (hier die Items innerhalb einer Vignette) zu einem Summenscore auf (siehe Abbildung 2.4) (Cook, Dodd & Fitzpatrick, 1999). Die Testlets (hier: Vignetten) können dann als polytome (Super-)Items aufgefasst werden. Die Bildung von Summenscores kann das Problem lokaler Abhängigkeiten zwischen den Items dahingehend auflösen, da für die Schätzung polytomer Raschmodelle nur eine lokale, stochastische Unabhängigkeit zwischen den Testlets (hier: Vignetten) vorausgesetzt wird (Rosenbaum, 1988). Nachteilig an diesem Ansatz ist, dass durch die Bildung eines Summenscores verschiedene Ankreuzmuster innerhalb der Testlets bei der statistischen Analyse nicht beachtet werden können.

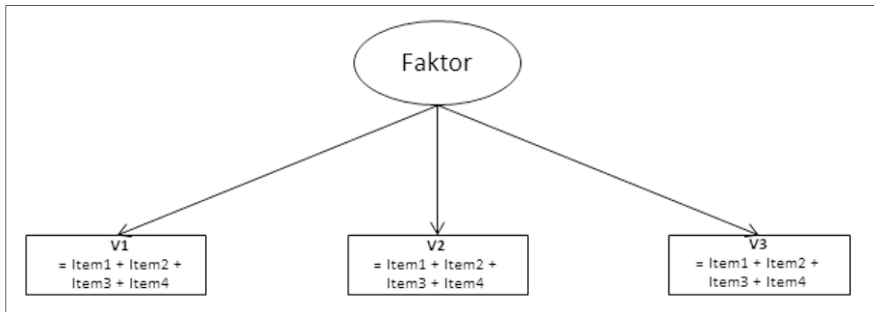


Abbildung 2.4: Schematische Darstellung für die Modellierung anhand des score-basierten Ansatzes für einen Vignettentest mit drei Vignetten (V1, V2, V3). Zu jeder Vignette wurden in diesem Beispiel vier geschlossene Items präsentiert.

Item-basierte Ansätze hingegen berücksichtigen die Testletstruktur auf Itemebene, beispielsweise durch die Spezifikation von Bi-Faktor-Modellen (Reise, 2012). Ein Vorteil bei der Verwendung von item-basierten Ansätzen ist, dass in die statistischen Analysen ein bzw. mehrere Parameter eingeführt werden, die die testlet-

spezifischen Effekte explizit berücksichtigen (Eckes, 2015). Im Falle von Vignettentests wird also für jede Vignette ein „Vignettenfaktor“ eingeführt, der die inhaltliche und statistische Abhängigkeit der Items, die durch den gemeinsamen Vignettenstamm verursacht wird, kontrollieren soll.

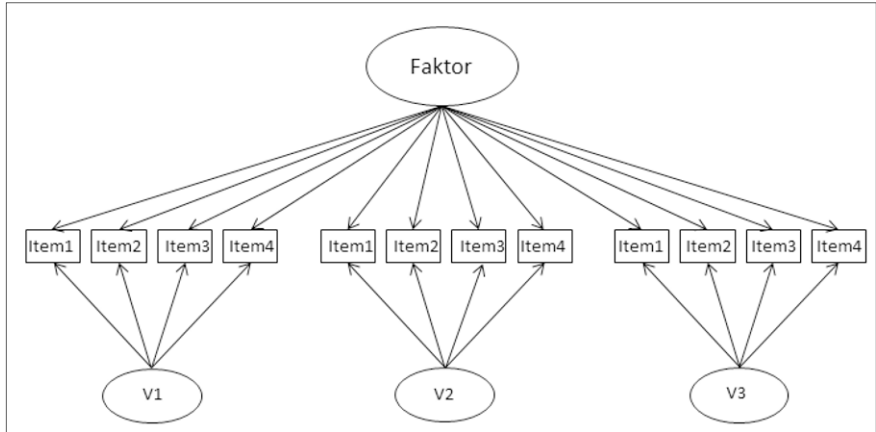


Abbildung 2.5: Schematische Darstellung für die Modellierung anhand des item-basierten Ansatzes durch die Anwendung eines Bi-Faktor Modells für einen Vignettentest mit drei Vignetten. Zu jeder Vignette wurden in diesem Beispiel vier geschlossene Items präsentiert. Die „Vignettenfaktoren“ V1, V2 und V3 sollen die Abhängigkeiten zwischen den Items kontrollieren.

Daneben werden auch Modelle vorgeschlagen, die die korrelativen Zusammenhänge zwischen den Items innerhalb von Testlets modellieren (Robitzsch & Lüdtke, 2014, 2015); diese werden aus Gründen der Übersichtlichkeit in diesem Artikel allerdings nicht weiterführend behandelt.

Es wurde beschrieben, dass es sowohl verschiedene Ansätze zur Punktvergabe von Vignetten (Itemebene vs. Relationenebene) als auch zur Modellierung (score-basiert vs. item-basiert) gibt. Der vorliegende Beitrag untersucht die Fragestel-

lung, ob diese verschiedenen Modellierungsansätze für vignettenbasierte Testverfahren als geeignet erscheinen. Als Kriterium wird die Abwesenheit von lokaler, stochastischer Abhängigkeit herangezogen.

Es wird erwartet, dass für eine Punktvorgabe auf Itemebene (aus methodischer Sicht) sowohl die Verwendung eines score-basierten als auch eines item-basierten Ansatzes geeignet ist, um lokale stochastische Abhängigkeiten zwischen den Items von Unterrichtsvignetten zu vermeiden. Für eine Punktvorgabe anhand von Itemrelationen kann hingegen angenommen werden, dass nur der score-basierte Ansatz zur Vermeidung von lokal stochastischer Abhängigkeit führt. Bei der Punktvorgabe anhand von Itemrelationen werden die Informationen aus zwei Items „zusammengezogen“; nicht die einzelnen Items werden bewertet, sondern Pseudoitems, die sich aus dem Vergleich des Antwortverhältnisses zweier (Einzel-)Items ergeben. Es ist daher anzunehmen, dass die Einführung von „Vignettenfaktoren“ wie im Bi-Faktor-Modell die Abhängigkeiten zwischen den Items, die durch die Relationenbildung entstehen, nicht aufheben kann.

2.2 Methode

Um die oben formulierte Fragestellung zu untersuchen, wurde ein vignettenbasiertes Testinstrument mit geschlossenem Antwortformat für den Bereich der Lesedidaktik herangezogen (Rutsch & Dörfler, in Druck). Dieses Testverfahren besteht aus zwölf Unterrichtsvignetten mit jeweils 4 bis 6 Items. Für diesen Test wurden zwei unterschiedliche Arten der Punktvorgabe vorgenommen.

Die Punktvorgabe auf Itemebene wurde so realisiert, dass bei exakter Übereinstimmung der Probandenbewertung mit dem aggregierten Expertenurteil ein Punkt vergeben wurde. Bei einer Abweichung der Probandenbewertung von der Expertenbewertung von $|x| = 1$ wurden 0.5 Punkte vergeben. Bei einer Abweichung der Probandenbewertung von der Expertenbewertung von $|x| > 1$ wurden 0 Punkte vergeben.

Für die Punktvergabe anhand von Itemrelationen wurden für die Unterrichtsvignetten jeweils 6 (bei 4 Items) bis 15 Itemrelationen (bei 6 Items) gebildet. Dabei wurde jedes Item mit jedem Item verglichen. Stimmte die gebildete Itemrelation des Probanden mit der Itemrelation des aggregierten Experten überein, erhielt der Proband 1 Punkt, ansonsten 0 Punkte (siehe auch Rutsch & Dörfler, in Druck). Es wurden nun die beiden Ansätze (score-basiert sowie item-basiert) hinsichtlich des Vorhandenseins von lokaler, stochastischer Abhängigkeit verglichen:

Für den score-basierten Ansatz wurden Vignettensummenscores gebildet: D. h. für die Punktvergabe auf Itemebene wurden die erreichten Punktzahlen pro Item (0, 0.5 oder 1 Punkt) summiert, sodass bei z. B. fünf Items für eine Vignette 0 bis 5 Punkte erreicht werden konnten³. Für die Punktvergabe anhand von Itemrelationen wurden die korrekt gebildeten Itemrelationen summiert⁴, sodass bei einer Vignette mit fünf Items 0 bis 10 Punkte erzielt werden konnten.

Für den item-basierten Ansatz wurde ein Bi-Faktor-Modell spezifiziert (Meschede et al., 2015; Reise, 2012). Dabei wird für jede Vignette ein eigener „Vignettenfaktor“ eingeführt, der die inhaltliche und statistische Abhängigkeit zwischen den Items einer Vignette bindet. Die Vignettenfaktoren der verschiedenen Vignetten werden im Modell als unkorreliert spezifiziert (vgl. Abbildung 2.5).

Um zu untersuchen, ob für die Verwendung der beiden Punktvergabesysteme der score-basierte bzw. der item-basierte Ansatz als geeigneter erscheint, wurden jeweils die lokalen, stochastischen Abhängigkeiten zwischen den Items untersucht. Dazu wurde die Q3-Statistik herangezogen (Yen, 1984; Yen, 1993), die im *R*-Package *mirt* implementiert ist (Chalmers, 2015). Die Q3-Statistik wird jeweils paarweise für die vorliegenden Testitems berechnet und stellt die Korrelation der Residuen dieser Items dar. Werte der Q3-Statistik von $> |0.2|$ weisen dabei auf

³ Die Vignetten im Testinstrument haben bis auf zwei Ausnahmen 5 Items. Für die Vignette mit 4 Items können die Probanden maximal 4 Punkte erreichen, für die Vignette mit 6 Items können die Probanden maximal 6 Punkte erreichen.

⁴ Für die Vignette mit 4 Items können die Probanden maximal 6 Punkte erzielen, da hier 6 Itemrelationen gebildet werden können; für die Vignette mit 6 Items können maximal 15 Punkte erzielt werden, da hier 15 Itemrelationen gebildet werden können.

ein Vorhandensein lokaler stochastischer Abhängigkeiten zwischen zwei Testaufgaben hin (Chen & Thissen, 1997; Yen, 1993).

Für die beiden Punktvergabesysteme wurde untersucht, inwiefern die Verwendung des score-basierten Ansatzes bzw. des item-basierten Ansatzes zu lokalen Abhängigkeiten zwischen den Items führt.

Stichprobe

Für die statistischen Analysen wurden die Daten von 764 Studierenden herangezogen (Alter: $M = 22.85$, $SD = 3.22$, 84.3 % weiblich). Die Stichprobe setzt sich aus Studierenden verschiedener Studiengänge zusammen: 40.7 % studieren Primarstufenlehramt mit dem Fach Deutsch, 18.7 % studieren Sekundarschullehramt mit dem Fach Deutsch (nicht gymnasiales Lehramt), 10.1 % studieren Sonderpädagogik, 6.2 % studieren gymnasiales Lehramt mit dem Fach Deutsch, 9 % studieren Lehramt ohne das Fach Deutsch, 8.8 % studieren Psychologie und 4.3 % studieren Germanistik im Bachelor-Studiengang.

2.3 Ergebnisse

Im Folgenden werden die Ergebnisse der Q3-Statistik zum Vorhandensein lokaler stochastischer Abhängigkeiten zwischen den Items für die beiden Punktvergabesysteme präsentiert.

Für die Punktvergabe auf Itemebene ist anhand von Abbildung 2.6 abzulesen, dass bei der Spezifikation eines Bi-Faktor-Modells keine lokalen Abhängigkeiten zwischen den einzelnen Items auftreten (Werte zwischen -0.2 und 0.2).

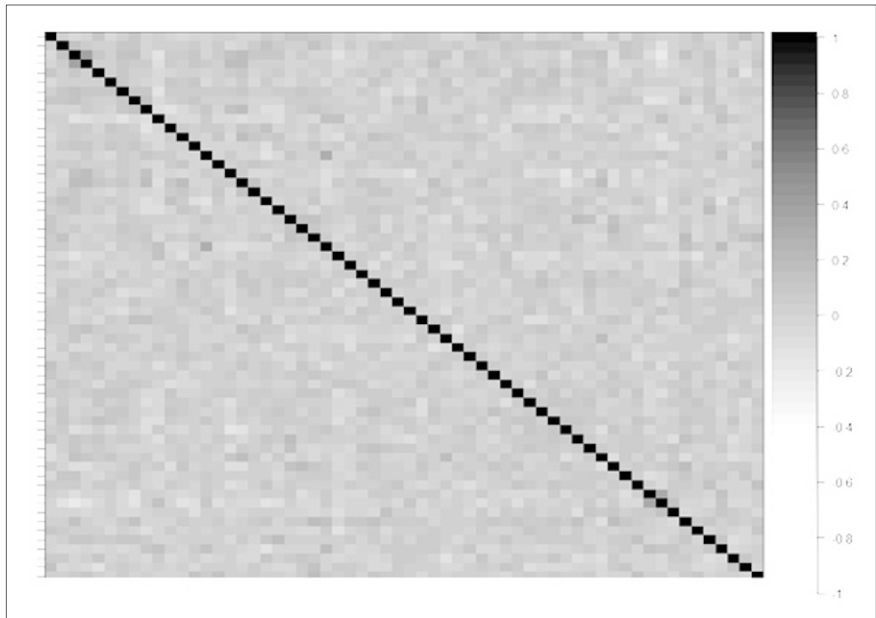


Abbildung 2.6: Darstellung der Q3-Statistik für die Punktvergabe auf Itemebene unter Verwendung eines item-basierten Ansatzes. Sehr helle bzw. sehr dunkle Felder zeigen Werte von < -0.2 bzw. > 0.2 an und indizieren lokale, stochastische Abhängigkeiten zwischen den Items.

Ebenso liegen bei der Bildung von Vignettensummenscores unter der Verwendung des score-basierten Ansatzes keine lokalen Abhängigkeiten zwischen den Testlets (hier Vignetten) vor (Werte zwischen -0.2 und 0.1). Abbildung 2.7 verdeutlicht dies.

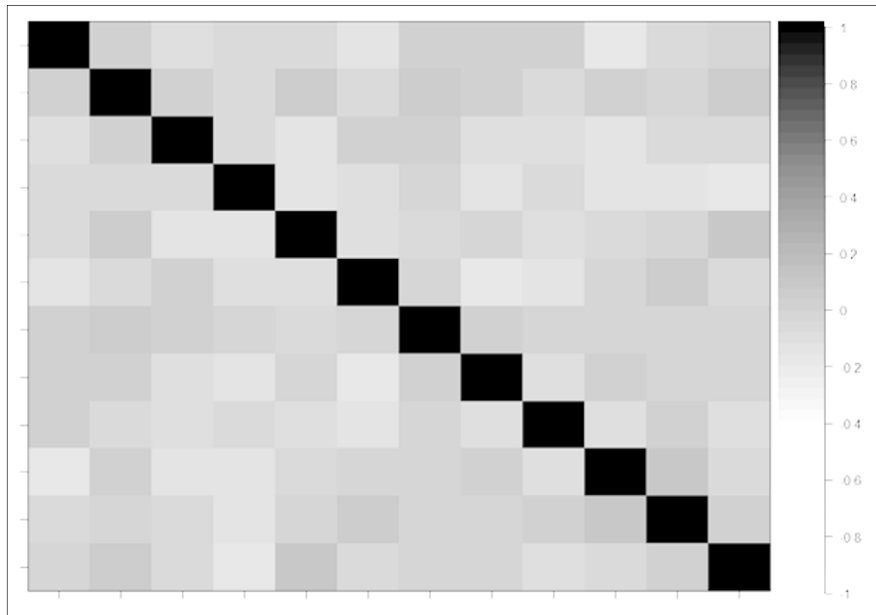


Abbildung 2.7: Darstellung der Q3-Statistik für die Punktwerte auf Itemebene unter Verwendung eines score-basierten Ansatzes. Sehr helle bzw. sehr dunkle Felder zeigen Werte von < -0.2 bzw. > 0.2 an und indizieren lokale, stochastische Abhängigkeiten zwischen den Items.

Für die Punktwerte anhand von Itemrelationen ergibt sich ein anderes Bild: Für die Anwendung eines Bi-Faktor-Modells zeigen sich bei der Berechnung der Q3-Statistik deutlich lokale Abhängigkeiten zwischen den Items (Werte zwischen -0.4 und 0.6): Es ergeben sich hier zwölf farblich gekennzeichnete Cluster, die die lokalen, stochastischen Abhängigkeiten der Itemrelationen innerhalb der zwölf Vignetten des Testinstruments repräsentieren (siehe Abbildung 2.8).

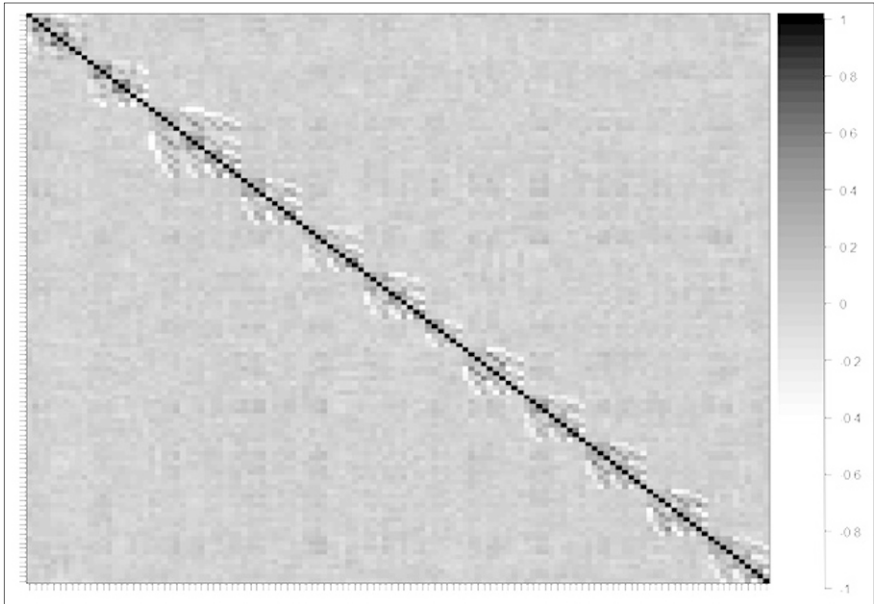


Abbildung 2.8: Darstellung der Q3-Statistik für die Punktvorgabe anhand von Itemrelationen unter Verwendung eines item-basierten Ansatzes. Sehr helle bzw. sehr dunkle Felder zeigen Werte von < -0.2 bzw. > 0.2 an und indizieren lokale, stochastische Abhängigkeiten zwischen den Items.

Für die Bildung von Vignettensummenscores ergeben sich für eine Punktvorgabe anhand von Itemrelationen keine lokalen Abhängigkeiten, wie Abbildung 2.9 zu entnehmen ist (Werte zwischen -0.2 und 0.1).

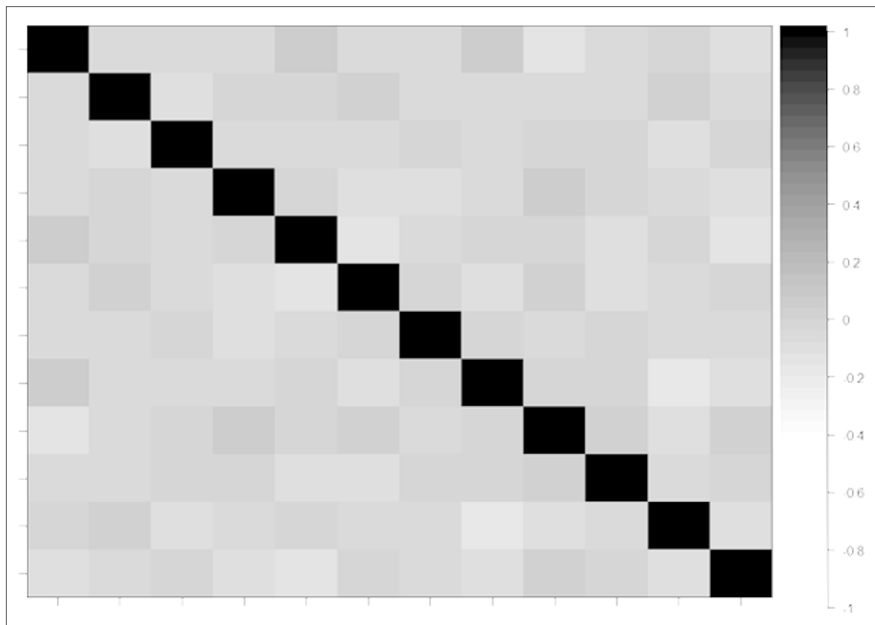


Abbildung 2.9: Darstellung der Q3-Statistik für die Punktvergabe anhand von Itemrelationen unter Verwendung eines score-basierten Ansatzes. Sehr helle bzw. sehr dunkle Felder zeigen Werte von < -0.2 bzw. > 0.2 an und indizieren lokale, stochastische Abhängigkeiten zwischen den Items.

2.4 Diskussion

Der vorliegende Beitrag thematisiert verschiedene methodische Herausforderungen, die sich aus der Verwendung von Unterrichtsvignetten mit geschlossenem Antwortformat als Testaufgaben zur Erfassung professioneller (Lehrer-)Kompetenzen ergeben. So stehen zur Bewertung der Unterrichtsvignetten anhand einer Expertennorm die Möglichkeiten zur Punktvergabe auf Itemebene sowie zur Punktvergabe anhand von Itemrelationen zur Verfügung. Darüber hinaus werden

in der Literatur score- und item-basierte Ansätze zur Modellierung von Testinstrumenten, die – wie Vignettentests – eine Testletstruktur aufweisen, vorgeschlagen. Das Ziel dieses Beitrages war es, anhand eines Vignettentests mit geschlossenem Antwortformat für den Bereich der Lesedidaktik (Rutsch & Dörfler, in Druck) zu untersuchen, ob unterschiedliche Arten der Punktvergabe jeweils unterschiedliche Möglichkeiten der statistischen Modellierung nahelegen. Als Beurteilungskriterium hierfür wurde die Abwesenheit von lokaler, stochastischer Abhängigkeit zwischen den Items bzw. den Itemrelationen innerhalb der Vignetten herangezogen. Das Vorhandensein von lokaler, stochastischer Unabhängigkeit stellt eine wichtige Voraussetzung für die Modellierung von latenten Merkmalen, beispielsweise mithilfe von Item-Response-Modellen, dar (Strobel, 2012). Item-Response-Modelle zur Erfassung latenter Merkmale, wie beispielsweise professioneller Kompetenzen, setzen die lokale, stochastische Unabhängigkeit der eingesetzten Testitems einerseits voraus, andererseits kann das Vorliegen lokaler, stochastischer Abhängigkeit zwischen den Testitems als Zielkriterium jedes Testinstruments verstanden werden (Rost, 1999).

Zur Indikation lokaler, stochastischer Abhängigkeiten zwischen den Items bzw. den Itemrelationen wurde die Q3-Statistik herangezogen (Yen, 1984; Yen, 1993). Die Erwartung, dass für eine Punktvergabe auf Itemebene aus methodischer Sicht sowohl ein score-basierter als auch ein item-basierter Ansatz zur Modellierung angemessen erscheint, wurde bestätigt. Für beide Modellierungsansätze sind keine lokalen, stochastischen Abhängigkeiten zwischen den Items zu beobachten. Dies wird durch die graphische Darstellung der Q3-Statistiken in Abbildung 2.6 und 2.7 ersichtlich.

Für eine Punktvergabe anhand von Itemrelationen hingegen erscheint ein score-basierter Ansatz überlegen zu sein, da nur so lokale Abhängigkeiten zwischen den Itemrelationen vermieden werden können. So wird in Abbildung 2.8 klar erkennbar, dass sich zwölf dunkel gefärbte Cluster bilden, die die lokalen, stochastischen Abhängigkeiten der Itemrelationen innerhalb der zwölf im Testinstrument verwendeten Vignetten repräsentieren. Die Einführung von „Vignettenfaktoren“ durch die Spezifikation eines Bi-Faktor-Modells kann in dieser Analyse

nicht zu einer Kontrolle der lokalen, stochastischen Abhängigkeiten zwischen den Itemrelationen beitragen. Für die Punktvorgabe anhand von Itemrelationen und Verwendung des score-basierten Ansatzes ergeben sich jedoch keine lokalen Abhängigkeiten (siehe Abbildung 2.9). Dieser Beitrag liefert daher erste Hinweise dafür, dass für eine Punktvorgabe anhand von Itemrelationen für Vignettentests mit geschlossenem Antwortformat ein score-basierter Ansatz verwendet werden sollte.

Für diesen Beitrag sind allerdings einige Einschränkungen zu beachten: Bei den hier berichteten Analysen handelt es sich (bisher) nur um einen einzelnen Vignettentest für den Bereich der Lesedidaktik. Für nachfolgende Analysen erscheint es lohnenswert, mehrere Testverfahren aus verschiedenen Domänen auf dieselbe Art und Weise zu untersuchen, sodass die hier erzielten Ergebnisse weiterführend untermauert werden können. Hierbei eignet sich das Forschungsprojekt EKoL in besonderem Maße, da hier vignettenbasierte Testverfahren mit geschlossenem Antwortformat aus verschiedenen Domänen (Deutsch, Geschichte, Mathematik, Naturwissenschaften und Technik) vorliegen.

Daneben ist anzumerken, dass in diesem Beitrag das Bi-Faktor-Modell als ein Beispiel für einen item-basierten Ansatz herausgegriffen wurde; es gibt allerdings noch weitere item-basierte Ansätze zur statistischen Modellierung von Testverfahren mit Testletstruktur (Eckes, 2015), die in nachfolgenden Untersuchungen ebenfalls eingesetzt und untersucht werden sollten.

Eine fundierte Beurteilung, welcher Ansatz zu welcher Punktvorgabe bzw. welcher Ansatz zur statistischen Modellierung für vignettenbasierte Testverfahren mit geschlossenem Antwortformat grundsätzlich geeignet ist, kann aufgrund der hier präsentierten Analysen (noch) nicht getroffen werden; dieser Beitrag stellt viel mehr einen ersten Anlauf der inhaltlichen sowie methodischen Auseinandersetzung mit diesem Thema dar. Abschließend kann diese Forschungsfrage erst beantwortet werden, wenn gesicherte Befunde darüber vorliegen, wie sich die (verschiedenen) Testwerte von vignettenbasierten Testverfahren mit geschlossenem Antwortformat, die sich durch verschiedene Punktvorgabearten und/oder verschiedene Modellierungsansätze ergeben, hinsichtlich Validität und Reliabilität

unterscheiden. Dies kann in nachfolgenden Analysen z. B. durch die Korrelation der erzielten Testwerte mit relevanten Außenkriterien festgestellt werden.

Darüber hinaus sollte neben einer methodischen Diskussion zur Modellierung verschiedener Punktvergabesysteme eine inhaltliche Diskussion über die *Bedeutung* der verschiedenen Punktvergabesysteme für Unterrichtsvignetten mit geschlossenem Antwortformat für die Erfassung professioneller Kompetenzen angestoßen werden. So sollte diskutiert werden, worin der inhaltliche Unterschied zwischen den beiden Punktvergabesystemen besteht. Diese Untersuchungen werden Gegenstand weiterführender Analysen im Forschungsprojekt EKoL sein.

Literatur

- Artelt, C., Beinicke, A., Schlagmüller, M. & Schneider, W. (2009). Diagnose von Strategiewissen beim Textverstehen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 41 (2), 96-103.
- Baer, M. & Buholzer, A. (2005). Beiträge über Forschung und Evaluation in der Ausbildung von Lehrkräften: Analyse der Wirksamkeit der berufsfeldorientierten Ausbildung für den Erwerb von Unterrichts- und Diagnosekompetenz. *Beiträge zur Lehrerbildung*, 23 (2), 243-248.
- Blömeke, S., König, J., Suhl, U., Hoth, J. & Döhrmann, M. (2015). Wie situationsbezogen ist die Kompetenz von Lehrkräften? *Zeitschrift für Pädagogik*, 61 (3), 310-327.
- Brovelli, D., Bölsterli, K., Rehm, M. & Wilhelm, M. (2013). Erfassen professioneller Kompetenzen für den naturwissenschaftlichen Unterricht - ein Vignettentest mit authentisch komplexen Unterrichtssituationen und offenem Antwortformat. *Unterrichtswissenschaft*, 41, 306-329.
- Brovelli, D., Bölsterli, K., Rehm, M. & Wilhelm, M. (2014). Using vignette testing to measure student science teachers' professional competencies. *American Journal of Educational Research*, 2 (7), 555-558.
- Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (Psychologie, 3. aktualisierte und erw. Aufl.). München: Pearson Studium.
- Chalmers, P. (2015). *Package „mirt“*. Verfügbar unter <https://cran.r-project.org/web/packages/mirt/mirt.pdf> [21.01.2016].

- Chen, W.-H. & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22 (3), 265-289.
- Cook, K. F., Dodd, B. G. & Fitzpatrick, S. J. (1999). A comparison of three polytomous item response theory models in the context of testletsScoring. *Journal of Outcome Measurement*, 3 (1), 1-20.
- Eckes, T. (2015). Lokale Abhängigkeit von Items im TestDaF-Leseverstehen. *Diagnostica*, 61 (2), 93-106.
- Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah NJ : Lawrence Erlbaum Associates.
- Massof, R. W. (2011). Understanding Rasch and item response theory models: Applications to the estimation and validation of interval latent trait measures from responses to rating scale questionnaires. *Ophthalmic Epidemiology*, 18 (1), 1-19.
- Meschede, N., Steffensky, M., Wolters, M. & Möller, K. (2015). Professionelle Wahrnehmung der Lernunterstützung im naturwissenschaftlichen Grundschulunterricht: Theoretische Beschreibung und empirische Erfassung. *Unterrichtswissenschaft*, 43 (4), 317-335.
- Oser, F. & Forster-Heinzer, S. (2015). Wer setzt das Maß? Eine kritische Auseinandersetzung mit dem Advokatorischen Ansatz. *Zeitschrift für Pädagogik*, 61 (3), 361-377.
- Oser, F., Heinzer, S. & Salzmann, P. (2010). Die Messung der Qualität von professionellen Kompetenzprofilen von Lehrpersonen mit Hilfe der Einschätzung von Filmvignetten. *Unterrichtswissenschaft*, 38 (1), 5-28.
- Pissarek, M. & Schilcher, A. (2015). Fachspezifische Lehrerkompetenzen im Fach Deutsch messen? Modellierung und Konstruktvalidierung eines Erhebungsinstruments im Rahmen der der Projektgruppe FALKO Regensburg. In C. R. Bräuer (Hrsg.), *Lehrende im Blick: Empirische Lehrerforschung in der Deutschdidaktik* (SpringerLink: Bücher, S. 321-342). Wiesbaden: Springer VS.
- Rehm, M. & Bölsterli, K. (2014). Entwicklung von Unterrichtsvignetten. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (S. 213-225). Berlin: Springer.
- Reise, S. P. (2012). Invited paper: The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, 47 (5), 667-696.
- Robitzsch, A. & Lüdtke, O. (2014). Zur (Nicht-)Modellierung lokaler Abhängigkeiten in Messmodellen: Weshalb der Modellfit kein geeignetes Kriterium für die Modellwahl ist. Verfügbar unter <https://sites.google.com/site/alexander/robitzsch/> [13.12.2016].

- Robitzsch, A. & Lüdtke, O. (2015). Kommentar zum Beitrag „Lokale Abhängigkeit von Items im TestDaF-Leseverstehen“ von Thomas Eckes. *Diagnostica*, 61 (2), 107-109.
- Rosenbaum, P. R. (1988). Items bundles. *Psychometrika*, 53 (3), 349-359.
- Rost, J. (1999). Was ist aus dem Rasch-Modell geworden? *Psychologische Rundschau*, 50 (3), 140-156.
- Rutsch, J. & Dörfler, T. (in Überarbeitung). Vignettentest zur Erfassung des fachdidaktischen Wissens im Leseunterricht bei angehenden Lehrkräften.
- Shavelson, R. J. (2013). On an approach to testing and modeling competence. *Educational Psychologist*, 48 (2), 73-86.
- Strobel, C. (2012). *Das Rasch-Modell: Eine verständliche Einführung für Studium und Praxis* (2. Aufl.). München: Rainer Hampp Verlag.
- Tepner, O. & Dollny, S. (2014). Entwicklung eines Testverfahrens zur Analyse fachdidaktischen Wissens. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (S. 311-323). Berlin: Springer.
- Wainer, H. & Kiely, G. (1987). Item clusters and computized adaptive testing: A case for testlets. *Journal of Educational Measurement*, 37 (3), 185-201.
- Wainer, H. & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement*, 37 (3), 203-220.
- Wilson, M. & Adams, R. J. (1995). Rasch models for item bundles. *Psychometrika*, 60 (2), 181-198.
- Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8 (2), 125-145. Verfügbar unter <http://conservancy.umn.edu/bitstream/handle/11299/107543/1/v08n2p125.pdf> [08.11.2016].
- Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30 (3), 187-213.

Effektive Kompetenzdiagnose in der Lehrerbildung
Professionalisierungsprozesse angehender Lehrkräfte
untersuchen

Rutsch, J.; Rehm, M.; Vogel, M.; Seidenfuß, M.; Dörfler, T.
(Hrsg.)

2018, VIII, 185 S. 29 Abb., Softcover

ISBN: 978-3-658-20120-3