

Für die testwissenschaftliche Begleitung der Erprobungsphase beauftragte die PASSAGE gemeinnützige Gesellschaft für Arbeit und Integration mbH Frau Dr. Almut Schön, TU Berlin, Zentraleinrichtung Moderne Sprachen.

Auszug aus dem Bericht

Pilotprojekt Fachsprachprüfungen für Pflegekräfte B2, 09/2020 – 05/2021, Erprobungsstandorte Hamburg und Bremen

1. Testgütekriterien
2. Bewertungskriterien
3. Der C-Test als Außenkriterium

1. Testgütekriterien

In der Testforschung sind die Hauptgütekriterien Objektivität, Reliabilität und Validität (s. u. a. Lienert & Raatz 1998; Bachman & Palmer 1996), weiterhin gibt es verschiedene Nebengütekriterien, auf die aber hier aus Platzmangel nicht weiter eingegangen werden kann.

Eine Prüfung ist objektiv, wenn die Ergebnisse nicht vom Prüfer oder anderen Prüfungsumständen abhängen, sondern tatsächlich nur von der Kompetenz und Performanz des Prüflings. Hier wird zwischen Durchführungs-, Auswertungs- und Interpretationsobjektivität unterschieden. Mit der Entwicklung des Leitfadens und einer Schulung für alle Prüfer ist in der Pilotierung der Fachsprachprüfungen für Pflege eine wichtige Basis für eine objektive Prüfungsdurchführung geschaffen worden. Im Echtbetrieb sind regelmäßige Prüferschulungen unverzichtbar, um standardisierte Prüfungsabläufe zu erreichen. In der Pilotierung erhielten alle Kandidaten dieselben Arbeitsanweisungen, hatten identische Bearbeitungszeiten und wurden mit ähnlichem Prüferverhalten konfrontiert, es kann also von einer hohen Durchführungsobjektivität in den Pilotprüfungen ausgegangen werden.

Die Fachsprachprüfung besteht aus zwei mündlichen Teilprüfungen (Gespräch mit einem Patienten und Gespräch mit einem Kollegen oder einem Arzt). Generell ist es nahezu unmöglich, Prüfungen, die in Gesprächsform stattfinden, vollkommen objektiv zu gestalten. Um dennoch eine hohe Auswertungs- und Interpretationsobjektivität zu erzielen, wurde ein detailliertes Bewertungsraster erarbeitet und in einer Schulung den Prüfern erklärt. Für den Echtbetrieb sind weitere und ausführlichere Schulungen im Umgang mit dem Bewertungsraster notwendig, denn nur so kann die Objektivität dieser vorwiegend mündlichen Prüfung erhöht werden.

Die Reliabilität gibt an, wie zuverlässig die Prüfung misst. Hierzu wird zumeist Cronbachs α berechnet. An den Pilotprüfungen nahmen 27 Prüflinge teil, dies ist aus statistischer Sicht eine geringe Zahl an Probanden, dennoch lassen sich anhand der Prüfungsbögen einige grundlegende statistische

Auswertungen vornehmen. Von 27 Prüflingen haben 11 die Prüfung bestanden, der Mittelwert betrug 53 von 90 Punkten, die Standardabweichung war 16,95 Punkte, das Minimum lag bei 25/90 Punkten, das Maximum bei 88/90 Punkten.

Mit den detaillierten Ergebnissen der Pilotprüfung wurde Cronbachs α berechnet, es ist mit $\alpha = 0,97$ extrem hoch, d. h. die Prüfung hat sehr zuverlässig gemessen. Auch durch Auslassung von Items wurde α kaum geringer. Schwierigstes Item war 3.4 „Kohärenz“ (im schriftlichen Teil), leichteste Aufgabe war 1.4 „Flüssigkeit“. Die Bewertungskriterien sind alle trennscharf. Einige Bewertungsskalen korrelieren sehr hoch miteinander ($> 0,8$), hier sollte bei weiteren Erprobungen oder auch im Echt-Betrieb weiter überprüft werden, ob diese Skalen dasselbe messen, da sie im Pilotbetrieb stets sehr ähnliche Ergebnisse lieferten. Dabei handelt es sich um die Skalen 1.1 und 1.2 („Maßnahmen erläutern und begründen“ und „auf Fragen reagieren“) und um die Skalen 2.3 („die nächsten Schritt zusammenfassen“) und 2.2 („therapeutische Konzepte besprechen“) und 2.4 („Interaktion“). Auch die Skalen 3.5 („Wortschatz“) und 3.1 („Vollständigkeit“) und 3.2 („Angemessenheit“) sowie 3.6 („Korrektheit“) und 3.2 („Angemessenheit“) und 3.4 („Kohärenz“) korrelieren sehr hoch miteinander und scheinen dasselbe zu messen. Es ist aber auch denkbar, dass diese Effekte in Folge einer verbesserten und ausführlicheren Bewertererschulung verschwinden würden.

Die Validität gibt an, inwiefern durch eine Prüfung gemessen wird, was gemessen werden soll, in diesem Fall Sprachkenntnisse im Deutschen auf dem Niveau B2. Die hier vorgestellten Szenarios, die in Kooperation zwischen Deutsch als Fremdsprache und Pflegewissenschaft entwickelt wurden, schaffen ideale Voraussetzungen für eine valide Prüfung, weil es sich um realistische Simulationen von typischen Berufssituationen in der Pflege handelt. In 27 Prüfungen wurden 10 Szenarien erprobt, damit liegen pro Szenario allerdings zu wenige Daten vor, um die evtl. Schwierigkeitsunterschiede rechnerisch bestimmen zu können. Hierzu wäre eine größere Datensammlung anzustreben. Die Validität der Prüfung wurde auch durch den Bewertungsbogen hergestellt. Alle Deskriptoren im Bewertungsbogen für die Bewertung mit 3 Punkten stammen ebenso wie die meisten Skalen direkt aus dem Gemeinsamen Europäischen Referenzrahmen für Sprachen (2001).

Die hier vorgestellte Prüfung ist also objektiv mit gewissen Einschränkungen, die allerdings das Format mündliche Prüfung generell betreffen. Sie ist weiterhin hoch reliabel und auch valide.

2. Bewertungskriterien

Für die Fachsprachprüfung für Pflegekräfte wurde ein eigenes Bewertungsraster entwickelt, das in jedem Prüfungsteil eine sechsstufige Bewertung verschiedener inhaltlicher und sprachlicher Aspekte sicher stellt:

Teil 1: Patientengespräch	
1.1 Maßnahmen erläutern und begründen	Aufgabenerfüllung
1.2 Auf Fragen reagieren	Aufgabenerfüllung
1.3 Interaktion	sprachliche Kompetenz
1.4 Flüssigkeit	sprachliche Kompetenz
1.5 Korrektheit	sprachliche Kompetenz
1.6 Aussprache	sprachliche Kompetenz
Teil 2: Gespräch mit einem Berufsangehörigen	
2.1 Informationen über Patient*innen weitergeben	Aufgabenerfüllung
2.2 therapeutische Konzepte besprechen und Verbesserungsvorschläge machen	Aufgabenerfüllung
2.3 die nächsten Schritte zusammenfassen, die eigenen Entscheidungen begründen bzw. die Vorschläge des anderen einbeziehen können	Aufgabenerfüllung
2.4 Interaktion	sprachliche Kompetenz
2.5 Korrektheit	sprachliche Kompetenz
2.6 Wortschatz	sprachliche Kompetenz
Teil 3: Schriftstück	
3.1 Vollständigkeit	Aufgabenerfüllung
3.2 Angemessenheit	Aufgabenerfüllung
3.3 Relevante Informationen auswählen und zuordnen	Aufgabenerfüllung
3.4 Kohärenz	sprachliche Kompetenz
3.5 Wortschatz	sprachliche Kompetenz
3.6 Korrektheit	sprachliche Kompetenz

Jede Bewertungsskala ist vierschrittig und kann mit 5 Punkten (= Übererfüllung), 3 Punkten (= Niveau B2), 2 Punkten (weniger als B2) oder 1 Punkt (= mangelhaft) bewertet werden. Insgesamt ist dann die maximale Bewertung 18 x 5 Punkten = 90 Punkte. Wenn in allen Kategorien 3 Punkte erreicht werden, ist das Niveau B2 nachgewiesen, das entspricht 54 Punkten, die auch als Bestehensgrenze festgelegt wurden. Die hier abgebildeten sprachlichen Kompetenzen stammen direkt aus dem Gemeinsamen Europäischen Referenzrahmen für Sprachen, s. dort Kapitel 4.4 und Kapitel 5.2 (Trim et al. 2001, S. 62ff. und 109ff.).

Die Gesprächsführungskompetenz, hier als „Interaktion“ bezeichnet, und die korrekte Sprachverwendung sind zentrale Kompetenzen für den beruflichen Alltag, sie werden daher in den Teilen 1 und 2 bewertet. Flüssigkeit, Aussprache und Wortschatzumfang sind ebenfalls wichtige Parameter der Sprachbeherrschung, sie wurden aber aus Gründen der Effizienz auf die Teile 1 und 2 verteilt. Es erscheint hilfreich, den Wortschatzumfang gezielt im Gespräch mit einem Berufsangehörigen zu bewerten, da hier ein umfangreicheres Vokabular nötig sein wird. Aussprache und Flüssigkeit unterscheiden sich eher nicht von Gespräch zu Gespräch, sie werden daher nur einmal im Teil 1 bewertet.

Im Ergebnis der Erprobung auch des Bewertungsbogens werden folgende Veränderungen vorgeschlagen:

- 3.1: die erforderliche Menge an Wörtern wird in die Auswertung mit einbezogen
- 1.1 oder 1.2 können vermutlich entfallen, da sie sehr ähnlich messen (s. o.), hier wird vorgeschlagen, die Bewertungsskala 1.1 zu entfernen
- Aus demselben Grund wird vorgeschlagen, 2.2 zu entfernen
- Im Teil 3 kann die Skala 3.2 entfallen, da sie sehr ähnlich wie 3.3 und 3.4 misst.
- Durch diese Veränderungen wird die Bewertungsskala gekürzt, die Bewertung wird effektiver. Die Bestehensgrenze läge dann bei $15 \times 3 = 45$ Punkten.

In der Erprobung hat sich gezeigt, dass etliche Kandidaten durch sehr gute mündliche Leistungen eine eher unterdurchschnittliche schriftliche Leistung kompensieren können. Dies betrifft vor allem die Kandidaten, die sich nach dem ebenfalls durchgeführten Deutschtest unter dem geforderten Niveau befinden.

Einerseits hat der Gesetzgeber die Überbetonung von mündlichen Kompetenzen durch die genau definierte Prüfungsform (s. GMK-Beschluss) vorgegeben, möglicherweise weil Heilberufe vor allem Sprechberufe sind. Andererseits stehen mündliche und schriftliche Kompetenzen im Zusammenhang und Pflegekräfte haben auch vielfältige Dokumentationsaufgaben, ganz abgesehen von im Arbeitsleben üblichen Schriftstücken wie Emails oder Briefen. Ein stärkeres Gleichgewicht könnte durch eine Doppelung aller Punkte im schriftlichen Bereich erzielt werden. Diese Änderungsvorschläge seien hier noch einmal zusammengefasst:

Teil 1: Patientengespräch				
	5	3	2	1
1.2 Auf Fragen reagieren				
1.3 Interaktion				
1.4 Flüssigkeit				
1.5 Korrektheit				
1.6 Aussprache				
Teil 2: Gespräch mit einem Berufsangehörigen				
	5	3	2	1
2.1 Informationen über Patient*innen weitergeben				
2.3 die nächsten Schritte zusammenfassen, die eigenen Entscheidungen begründen bzw. die Vorschläge des anderen einbeziehen können				
2.4 Interaktion				
2.5 Korrektheit				
2.6 Wortschatz				
Teil 3: Schriftstück				
	10	6	4	2
3.1 Vollständigkeit				
3.3 Relevante Informationen auswählen und zuordnen				
3.4 Kohärenz				
3.5 Wortschatz				
3.6 Korrektheit				

Es wären somit maximal 100 Punkte erreichbar, die Bestehensgrenze B2 läge bei 90 Punkten.

3. Ergebnisse im Zusammenhang mit dem Einsatz eines unabhängigen Deutschtests

An den Pilotprüfungen nahmen 27 Pfleger und Pflegerinnen mit sehr unterschiedlichen sprachlichen Hintergründen teil, die Personen kamen aus 15 verschiedenen Herkunftsländern. Sie haben damit sehr unterschiedliche Deutschlernbiographien und Lernstände. Um die Ergebnisse dieser Pilotstudie abzusichern und auch um mehr Erkenntnisse über internationale Pflegekräfte zu gewinnen, wurden alle Prüflinge gebeten, zusätzlich zur Fachsprachenprüfung, einen zwanzigminütigen C-Test in Deutsch abzulegen. C-Tests haben sich als Messinstrument in zahlreichen Studien immer wieder bewährt, wenn es darum geht, Sprachkenntnisse global einzuschätzen und finden mittlerweile in vielen Sprachen ihre Anwendung (vgl. Grotjahn 2002)

Auch für diese Studie wurde ein C-Test eingesetzt, um angesichts der Heterogenität der Prüflinge und der Neuartigkeit der Fachsprachenprüfung für Pflegekräfte ein Außenkriterium zu haben. Der eingesetzte Test wurde an der ZEMS der TU Berlin entwickelt, pilotiert und kalibriert, s. Schön, Zimmermann, Johnson 2012.

Alle Prüflinge haben den C-Test abgelegt, das Minimum lag bei 15/100 Punkten, das Maximum bei 65/100 Punkten, der Mittelwert bei 39,70/100, die Standardabweichung betrug 10,3 Punkte. Laut C-Test verteilten sich die Sprachkenntnisse in dieser kleinen Gruppe wie folgt:

Niveau lt GER	Personen	in %
A1	1	3%
A2	8	30%
B1	16	60%
B2	2	7%
Summe	27	100

Es gab nach C-Test, aber auch nach Einschätzung der Bewerterinnen viele Prüflinge mit zu geringen Sprachkenntnissen, so erklärt sich auch die sehr hohe Durchfallquote.

Allerdings stimmen C-Test und Bewertung der Fachsprachenprüfung nicht immer überein. Beide Ergebnisse sind nach Shapiro-Wilk normal verteilt, es lässt sich daher die Korrelation nach Pearson berechnen. Diese beträgt $r = 0,32$ und befindet sich damit im mittleren Bereich, man kann von einem positiven, moderaten Zusammenhang zwischen C-Test-Ergebnis und dem Ergebnis der FSP sprechen. Würde man die Bewertungskriterien wie oben vorgeschlagen verändern, würde sich dieser Zusammenhang auf $r = 0,37$ verbessern.

Auch bei der Auswertung dieses Zusammenhangs fällt auf, dass zum Einen etliche Prüflinge die Prüfung bestanden haben, obwohl ihr C-Test-Ergebnis zu gering ist und dass diese Prüflinge zum Anderen vor allem durch gute mündliche Leistungen die eher mangelhaften schriftlichen Leistungen kompensieren konnten.

Insgesamt hat jedoch der Einsatz eines Außenkriteriums die Qualität der Fachsprachenprüfung bestätigt und erhärtet, dass hier tatsächlich valide die Deutschkompetenz im beruflichen Kontext geprüft wird.

Literatur:

Bachman, Lyle F.; Palmer, Adrian S. (1996): *Language testing in practice. Designing and developing useful language tests*. Oxford: Oxford University Press.

Grotjahn, Rüdiger (Hrsg.) (2002): *Der C-Test. Theoretische Grundlagen und praktische Anwendungen. (Bd. 4)*. Bochum: AKS-Verlag.

Lienert, Gustav A.; Raatz, Ulrich (1998): *Testaufbau und Testanalyse*. 6. Auflage. Weinheim: Beltz.

Schön, Almut; Zimmermann, Kerstin; Johnson, Natalia (2012): *Intrauniversitäre Kooperation – zur gemeinsamen Entwicklung eines C-Tests durch Sprachenzentrum und Sprachlehrforschung*. In: *Fremdsprachen und Hochschule* 86, S. 61–79.

Trim, John L. M.; North, Brian; Coste, Daniel (2009): *Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen. Niveau A1, A2, B1, B2, C1, C2*. Unter Mitarbeit von Marion Butz und Jürgen Quetz. 8. Aufl. Berlin, Wien u.a.: Langenscheidt.