Prüfungen/Psychometrie (Vorträge)

V1-612 (174)

Stabile Antwortmuster bei Script Concordance Test Fragen in der Schweizer Facharztprüfung Allgemeine Innere Medizin

Daniel Stricker[1], Felicitas-Maria Lahner[1], Raphael Bonvin[2], Christoph Berendonk[1]

[1]Bern, Schweiz
[2]Lausanne, Schweiz

Fragestellung: Mit dem Script Concordance Test (SCT) soll die Fähigkeit zu klinischem Denken (clinical reasoning) geprüft werden [1], [2]. Jedoch wird das Fragenformat u.a. kritisiert, weil die Messzuverlässigkeit schwierig zu überprüfen ist. Insbesondere fehlen Angaben zu test-retest Reliabilitäten [3]. Ziel der vorliegenden Studie ist es, die Stabilität der Antwortmuster auf SCT Fragen zu untersuchen.

Methoden: In zwei Facharztprüfungen für Allgemeine Innere Medizin in der Schweiz (Juni 2014, November 2014) wurden (neben 100 Multiple Choice Fragen) jeweils 20 SCT Fragen eingesetzt, zehn davon waren in beiden Prüfungen identisch. Insgesamt nahmen 591 Kandidaten an den Prüfungen Teil (Juni:287, November:304). Alle Kandidaten stammten aus derselben Weiterbildungskohorte und konnten frei zwischen beiden Terminen wählen. Die SCT Items wurden auf einer fünfstufigen Skala beantwortet und mit aggregate Scoring [4] basierend auf einem Expertenpanel (N=26) bewertet.

Wir verglichen die wiederholten SCT-Items bezüglich der Verteilung der Antworten der Kandidaten über die unterschiedlichen Antwortalternativen.

Ergebnisse: Die mittlere Leistung der beiden Prüfungsgruppen in den wiederholten SCT Items ist identisch (Juni: M=7.44, SD=0.975; November: M=7.45, SD=0.939). Hochgerechnet auf 100 Fragen beträgt das Cronbach-a des SCT Teils.77 (Juni) resp. 67 (Nov.). Der Vergleich der Antwortmuster der beiden Prüfungsgruppen war für alle 10 wiederholten SCT Fragen identisch.

Diskussion: Diese Resultate legen nahe, dass die Messzuverlässigkeit der wiederverwendeten SCT Fragen als hoch einzustufen ist. Auch wenn mit dieser Untersuchung längst nicht alle methodischen Probleme der SCT Fragen im summativen Einsatz geklärt werden [3] ist das Resultat dennoch bemerkenswert, weil beide Gruppen unabhängig waren und zwischen den beiden Erhebungen 5 Monate lagen.

Take home messages: Diese Studie zeigt, dass der sorgfältige Einsatz von SCT Fragen zu stabilen Resultaten führen kann.

Literatur

1. Charlin B, van der Vleuten C. Standardized Assessment of Reasoning in Contexts of Uncertainty: The Script Concordance Approach. Eval Health Prof. 2004;27(3):304-319. DOI: 10.1177/0163278704267043
2. Lubarsky S, Dory V, Duggan P, Gagnon R, Charlin B. Script concordance testing: from theory to practice: AMEE guide no. 75. Med Teach. 2013;35(3):184-193. DOI: 10.3109/0142159X.2013.760036
3. Lineberry M, Kreiter CD, Bordage G. Threats to validity in the use and interpretation of script concordance test scores. Med Educ. 2013;47(12):1175-1183. DOI: 10.1111/medu.12283
4. Bland AC, Kreiter CD, Gordon JA. The psychometric properties of five scoring methods applied to the script concordance test. Acad Med. 2005;80(4):395-399. DOI: 10.1097/00001888-200504000-00019

Bitte zitieren als: Stricker D, Lahner FM, Bonvin R, Berendonk C. Stabile Antwortmuster bei Script Concordance Test Fragen in der Schweizer Facharztprüfung Allgemeine Innere Medizin. In: Jahrestagung der Gesellschaft für Medizinische Ausbildung (GMA). Bern, 14.-17.09.2016. Düsseldorf: German Medical Science GMS Publishing House; 2016. DocV1-612.
DOI: 10.3205/16gma174, URN: urn:nbn:de:0183-16gma1745
Frei verfügbar unter: http://www.egms.de/en/meetings/gma2016/16gma174.shtml

V1-643 (175)

Psychometrische Gütekriterien von Multiple-Choice-Examen in Abhängigkeit der Anzahl Kandidaten und Items: Ab welchen Stichprobengrössen sind die Gütekriterien vertrauenswürdig?

Rainer Hofer, Sören Huwendiek

Bern, Schweiz

Fragestellung: In Prüfungsanalysen mit einer kleinen Anzahl von Kandidaten und/oder Items wird die Aussagekraft der psychometrischen Gütekriterien in der Itemanalyse respektive in der Interpretation zum Teil nicht ausreichend mitberücksichtigt. In der vorliegenden Studie wurde der Frage nachgegangen, ab welcher Anzahl Kandidaten und Items die Gütekriterien ausschliesslich in dem als vertrauenswürdig bestimmten Intervall liegen.

Methode: Der Studie lagen die Daten der Eidgenössischen Prüfung Humanmedizin des Jahres 2014 zugrunde, bei der 592 deutschsprachige Kandidaten 300 Multiple-Choice-Fragen beantworteten. Die Daten der 269 französischsprachigen Kandidaten wurden nicht berücksichtigt, um soziokulturelle Einflüsse bestmöglich auszuschliessen. Als Ausgangslage dienten die Werte der Gütekriterien (wie Reliabilität, Schwierigkeit, Standardmessfehler, Trennschärfe) über alle 592 Kandidaten und alle 300 Items. Für diese Werte wurden die 95%-Vertrauensintervalle bestimmt.

Mittels Bootstrapping [1] wurden danach 100 Stichproben aus der Grundgesamtheit gezogen. Über die 100 Ziehungen wurden die Gütekriterien gemittelt, deren Vertrauensintervalle berechnet und diese mit den Ausgangswerten verglichen.