

Das Bootstrapping wurde iterativ mit jeweils um 10 Kandidaten und/oder 10 Items reduzierten Stichproben wiederholt. In der entgegengesetzten Iteration wurden mit Stichproben von 10 Kandidaten und 10 Items gestartet und anschliessend die Stichproben nach dem Zufallsprinzip jeweils um 10 Kandidaten und/oder 10 Items vergrössert. Die Analyse wurde mit dem Statistikpaket R durchgeführt.

Ergebnisse: Die Daten werden bis zum Beitrag zu Ende ausgewertet sein. Es wird aufgezeigt werden, ab welchen Stichprobengrössen (Kandidaten, Items) die Gütekriterien „vertrauenswürdig“ sind.

Diskussion: Die Ergebnisse werden anhand der Literatur diskutiert werden.

Take-Home-Messages: Entsprechend der Ergebnisse werden Take Home Messages formuliert werden.

Literatur

1. Efron B. Bootstrapping methods: another look at the jackknife. *Ann Statist.* 1979;7:1–26. DOI: 10.1214/aos/1176344552

Bitte zitieren als: Hofer R, Huwendiek S. Psychometrische Gütekriterien von Multiple-Choice-Examen in Abhängigkeit der Anzahl Kandidaten und Items: Ab welchen Stichprobengrössen sind die Gütekriterien vertrauenswürdig? In: Jahrestagung der Gesellschaft für Medizinische Ausbildung (GMA). Bern, 14.-17.09.2016. Düsseldorf: German Medical Science GMS Publishing House; 2016. DocV1-643. DOI: 10.3205/16gma175, URN: urn:nbn:de:0183-16gma1757
Frei verfügbar unter: <http://www.egms.de/en/meetings/gma2016/16gma175.shtml>

V1-662 (176)

Einfluss von unterschiedlichen Bewertungs-Algorithmen für Kprim Fragen auf psychometrische Charakteristiken von Prüfungen

Felicitas-Maria Lahner¹, Zineb Nouns¹, Martin R. Fischer², Sören Huwendiek¹

¹Bern, Schweiz

²München, Deutschland

Fragestellung/Zielsetzung: Die Vor- und Nachteile unterschiedlicher Bewertungs-Algorithmen von Kprim-Fragen [1], [2], ist nicht eindeutig geklärt. Diese Studie untersucht den Einfluss verschiedener Bewertungs-Algorithmen für Kprim-Fragen auf deren psychometrische Parameter und vergleicht diese mit denen von Typ A-Fragen.

Methoden: Wir untersuchten an einer Stichprobe von 38 Prüfungen (998 Kprim und 2163 Typ A Items, durchschnittlich 225 Kandidaten/Prüfung) zweier Schweizer Fakultäten sowie der Eidgenössischen Prüfung den Einfluss unterschiedlicher Bewertungs-Algorithmen für Kprim Fragen auf Reliabilität, Trennschärfe, Schwierigkeit und die Gesamtpunktzahl.

Wir verglichen drei Bewertungs-Algorithmen für Kprim Items mit 4 Antwortmöglichkeiten:

1. Viertelpunkt-Bewertung (VP): für jede richtige Teilantwort $\frac{1}{4}$ Punkt
2. Halbpunkt-Bewertung (HP): $\frac{1}{2}$ Punkt wenn mehr als die Hälfte, 1 Punkt, wenn alle Teilantworten richtig beantwortet.
3. Ganzpunkt-Bewertung (GP): 1 Punkt wenn alle Teilantworten richtig beantwortet.

Zum Vergleich wurden Typ A Fragen miteinbezogen. Die Bewertungs-Algorithmen verglichen wir mit Varianzanalysen für wiederholte Messungen bzw. Friedmann Tests falls die Voraussetzungen für Varianzanalysen nicht erfüllt wurden.

Ergebnisse: VP und HP führen zu signifikant höheren Reliabilitäten und Trennschärfen im Vergleich zu GP und Typ A. Im Bezug auf die Itemschwierigkeit unterscheiden sich alle Bewertungs-Algorithmen signifikant, wobei VP leichteren und GP zu den schwierigeren Items führt. HP führt zu leichteren Items als Typ A. Bei der Gesamtpunktzahl zeigt sich, dass Kandidaten mit VP im Durchschnitt signifikant mehr Punkte erreichen als mit HP oder GP.

Diskussion: Bewertungs-Algorithmen mit Teilpunkten führen zu besseren psychometrischen Charakteristiken. Dies zeigt sich auch in anderen Studien zu Fragen mit Mehrfachantworten wie z.B. Pick-N [3] oder kleineren Studien mit Kprim-Fragen [4].

Take home message: Bewertungen mit Teilpunkten führen bei Kprim-Fragen zu besseren psychometrischen Charakteristiken.

Literatur

1. Javid L. The Comparison between Multiple-choice (MC) and Multiple True-false (MTF) Test Formats in Iranian Intermediate EFL Learners' Vocabulary Learning. *Procedia Soc Behav Sci.* 2014;98:784-788. DOI: 10.1016/j.sbspro.2014.03.482
2. Mobalegh A, Barati H. Multiple True-false (MTF) and Multiple-choice (MC) Test Formats: A Comparison between Two Versions of the Same Test Paper of Iranian NUÉE. *J Lang Teach Res.* 2012;3(5):1027-1037. DOI: 10.4304/jltr.3.5.1027-1037
3. Bauer D, Holzer M, Kopp V, Fischer MR. Pick-N multiple choice-exams: a comparison of scoring algorithms. *Adv Health Sci Educ Theory Pract.* 2011;16(2):211-221. DOI: 10.1007/s10459-010-9256-1
4. Albanese MA, Sabers DL. Multiple True-False Items: A Study of Interitem Correlations, Scoring Alternatives, and Reliability Estimation. *J Educ Meas.* 1988;25(2):111-123. DOI: 10.1111/j.1745-3984.1988.tb00296.x

Bitte zitieren als: Lahner FM, Nouns Z, Fischer MR, Huwendiek S. Einfluss von unterschiedlichen Bewertungs-Algorithmen für Kprim Fragen auf psychometrische Charakteristiken von Prüfungen. In: Jahrestagung der Gesellschaft für Medizinische Ausbildung (GMA). Bern, 14.-17.09.2016. Düsseldorf: German Medical Science GMS Publishing House; 2016. DocV1-662. DOI: 10.3205/16gma176, URN: urn:nbn:de:0183-16gma1768
Frei verfügbar unter: <http://www.egms.de/en/meetings/gma2016/16gma176.shtml>