Tibor Vígh

Qualitätskriterien in der Messung und Bewertung von Schreibfertigkeit

Die Ergebnisse einer Analyse mit dem Partial-Credit-Modell

1. Einleitung

Die Messung und Bewertung von Schreibfertigkeit ist ein wichtiges Gebiet von internationalen Sprachtestforschungen. Schreibfertigkeit wird in kommunikativen Sprachprüfungen mit Performanztests gemessen, deren Ergebnisse mit analytischen Bewertungsverfahren bestimmt werden. Das Hauptziel meiner empirischen Untersuchung war, Kriterien zu definieren, mit denen die Qualität dieser Verfahren untersucht und erhöht werden kann.

Im ersten Teil des vorliegenden Beitrags stelle ich den Hintergrund der empirischen Untersuchung vor. Zuerst wird ein Überblick darüber gegeben, welche Faktoren die Messung und Bewertung von Schreibfertigkeit bestimmen können. Nach der Darstellung der Forschungsfragen und Methoden werden die Ergebnisse präsentiert. In der empirischen Untersuchung wurden die Charakteristika der Bewertungskriterien, die Fähigkeit der Kandidaten und die Struktur der Ratingskalen im Kontext des Abiturs auf der Stufe B2 für Deutsch als Fremdsprache in Ungarn analysiert. Zum Schluss werden die Qualitätskriterien zusammengefasst und die wichtigsten Konsequenzen gezogen.

2. Faktoren der Messung und Bewertung von Schreibfertigkeit

Der Messwert von Schreibleistungen wird durch die Fähigkeit der Kandidaten, die Charakteristika der Bewerter, die Struktur der Bewertungskriterien und Ratingskalen beeinflusst (vgl. die Modelle von Engelhard 1992: 173 und Eckes

2005: 78). Die getrennte Analyse dieser Faktoren kann nur mit den Modellen der probabilistischen Testtheorie (auch Item-Response-Theorie genannt, abgekürzt: IRT) erfolgen. Um die Leistungen der Prüfungskandidaten zu modellieren, ist die Anwendung des Partial-Credit-Modells (Masters 1982) möglich. Zur Analyse der Ratereffekte kann das Multifacetten-Rasch-Modell von Linacre (1994) eingesetzt werden, das für die Skalierung von polytomen Daten ebenfalls geeignet ist. Im Folgenden wird die Verwendung dieser Modelle bei der Analyse von Testaufgaben und analytischen Bewertungsverfahren zu Schreibfertigkeit kurz vorgestellt.

Mit den Verfahren der klassischen Testtheorie kann die Wirkung von Hintergrundvariablen (z.B. die kognitiven und affektiven Merkmale der Testpersonen) analysiert werden, die die Schreibfertigkeit der Kandidaten bestimmen. Mit den IRT-Modellen für mehrstufige Daten können die Fähigkeitsniveaus der Probanden zahlenmäßig nachgewiesen und in einem Intervall charakterisiert werden. Außerdem kann festgestellt werden, wie gut die in den Ratingskalen definierten Werte diesem Intervall entsprechen (Park 2004).

Die Beurteiler können bei der Bewertung von Schülerleistungen, bei der Interpretation von Kriterien und beim Gebrauch von Ratingskalen zu unterschiedlichen Schlussfolgerungen kommen (vgl. Engelhard/Myford 2003, Eckes 2005, 2008). Aus diesem Grund sind die Prüferschulungen nur teilweise dazu geeignet, die eigenen Norminterpretationen der Beurteiler zu vereinheitlichen (vgl. Barrett 2001), deshalb sollten die Strenge bzw. die Milde der Bewerter untersucht werden. Dazu ist das Multifacetten-Rasch-Modell (Linacre 1994) verwendbar, mit dem die Homogenität der Beurteilergruppe analysiert werden kann (vgl. Eckes 2004).

Die Rolle der Bewertungskriterien ist entscheidend bei der Bestimmung des Messwerts. Anhand dieser Kriterien wird die schriftliche Textproduktion der Kandidaten beurteilt und mit Punktzahlen bewertet (vgl. Lumley 2002, Eckes 2008). Mit den IRT-Modellen für polytome Daten können die Schwierigkeit und die Modellpassung der Bewertungskriterien nachgewiesen werden (vgl. Engelhard/Myford 2003, Park 2004, Porsch 2010).

Die Bewertungsskalen funktionieren angemessen, wenn sie dazu geeignet sind, die Schreibfertigkeitsniveaus der Probanden genau zu beschreiben. Mit der Verwendung des Partial-Credit-Modells (Masters 1982) kann analysiert werden, wie gut die Bewerter die Ratingskalen verwenden, ob alle Deskriptoren in gleicher Weise gebraucht werden (Neumann 2007). So kann die empirische Validierung der Skalen gesichert werden (Park 2004).

Literaturverzeichnis

Barrett, S. (2001): The impact of training on rater variability, in: *International Education Journal*, 2, 1, 49–58.

Eckes, T. (2004): Beurteilerübereinstimmung und Beurteilerstrenge: Eine Multifacetten-Rasch-Analyse von Leistungsbeurteilungen im „Test Deutsch als Fremdsprache" (TestDaF), in: *Diagnostica*, 50, 2, 65–77.

Eckes, T. (2005): Evaluation von Beurteilungen: Pychometrische Qualitätssicherung mit dem Multifacetten-Rasch-Modell, in: *Zeitschrift für Psychologie*, 213, 2, 77–96.

Eckes, T (2008): Rater types in writing performance assessments: A classification approach to rater variability, in: *Language Testing*, 25, 2, 155–185.

Engelhard, G. (1992): The measurement of writing ability with a many-facet Rasch model, in: *Applied Measurement in Education*, 5, 3, 171–191.

Engelhard, G./Myford, C.M. (2003): *Monitoring faculty consultant performance in the Advanced Placement English Literature and Composition Program with a many-faceted Rasch model*, New York: College Entrance Examination Board.

Europarat (2001): *Gemeinsamer europäischer Referenzrahmen für Sprachen: Lernen, lehren, beurteilen*, Berlin: Langenscheidt.

Linacre, J.M. (1994): *Many-facet Rasch Measurement*, Chicago: MESA Press.

Lumley, T. (2002): Assessment criteria in a large-scale writing test: What do they really mean to the raters?, in: *Language Testing*, 19, 3, 246–276.

Masters, G.N. (1982): A Rasch model for partial credit scoring, in: *Psychometrika*, 47, 2, 149–174.

Neumann, A. (2007): *Briefe schreiben in Klasse 9. und 11. Beurteilungskriterien, Textstrukturen und Schülerleistungen*, Münster: Waxmann.

Park, T. (2004): An investigation of an ESL placement test of writing using many-facet Rasch measurement, in: *Working Papers in TESOL & Applied Linguistics*, 4, 1, 1–21.

Porsch, R. (2010): Die Erprobung eines Kodierschemas zur Messung der Schreibkompetenz im Fach Französisch, in: R. Porsch/B. Tesch/O. Köller (eds): *Standardbasierte Testentwicklung und Leistungsmessung. Französisch in der Sekundarstufe I,* Münster: Waxmann, 267–286.

Vígh, T. (2011): Qualitätskriterien der Messung und Bewertung von Schreibfertigkeit. Die Ergebnisse einer Analyse mit dem Partial-Credit-Modell, Universität Bremen. http://www.fremdsprachenzentrum-bremen.de/fileadmin/autor/dateien/Symposion_2011/ppt/ag2/Vigh.pdf (20.7.2011).

Wu, M./Adams, R.J./Wilson, M.R. (1998): *ACER ConQuest. Generalised Item Response Modelling Software*, Australia: ACER Press.

Dr. Tibor Vígh, Universität Szeged, Institut für Erziehungswissenschaft, Petőfi Sándor sgt. 30–34, 6722 Szeged, Ungarn, vigh.tibor@edpsy.u-szeged.hu