Relativität und Normativität in der unterrichtspraktischen und psychometrischen Leistungsbewertung

Samuel Merk & Sarah Bez

1. Einleitung

Der vorliegende Beitrag führt in die häufig zitierte Unterscheidung von „Bezugsnormen“ ein, um anhand dieser zu diskutieren, welche Rolle Relativität und Normativität bei der Leistungsbewertung in Bildungspraxis und -forschung spielen. Die zentrale These lautet dabei, dass sowohl die (empirisch quantitative) Bildungsforschung als auch die Bildungspraxis Bewertungen primär durch Relativierung einzelner Messwerte an einer Verteilung dieser Messwerte konstruieren, wobei Normativität bei der Wahl der betrachteten Größe (z.B. intraindividueller Leistungszuwachs; Abstand von einem Referenzwert) sowie der Transformation von Prozenträngen in Noten oder andere Bewertungsskalen (z.B. Notenpunkte) erzeugt wird.

2. Bezugsnormen

2.1. Einführendes Beispiel

Angenommen, eine Grundschullehrkraftlehrkraft führt in ihrer Klasse alle drei Wochen einen Test zum sinnkonstruierenden Satzlesen durch, in dem Schüler*innen unvollständige Sätze (z. B. *„Mama holt meinen ab.“*) durch die Auswahl eines Wortes aus einer Liste (z. B. *„Bruder – Kopf – Sinn – Schatten“*) sinnvoll ergänzen müssen. Die Schüler*innen haben drei Minuten Zeit, um möglichst viele (der gleich schwierigen) Sätze zu bearbeiten (Speedtest; Rost, 2004). Abschließend wird gezählt, wie viele der Aufgaben korrekt bearbeitet wurden. Angenommen, der Lehrkraft liegen die folgenden Ergebnisse ihrer Klasse vor (siehe Tabelle 1). Wie sind die Ergebnisse des letzten Tests zu bewerten? Was macht eine „gute“ Leistung in diesem Fall aus?

Tabelle 1: Fiktive Ergebnisse eines Leseverständnistests. W1, W4, etc. entsprechen Woche 1, Woche 4, etc.

	Woche				
Schüler*in	**1**	**4**	**7**	**10**	**13**
A	5	7	7	9	10
B	2	4	4	7	11
C	10	9	11	12	13
D	4	5	6	7	6
E	1	2	1	3	3
F	4	3	4	5	8

Legt man Lehrkräften eine solche Aufgabe vor, sind vor allem drei Argumentationsstränge zu beobachten (Rheinberg, 1980): Zum einen wird argumentiert, dass Schüler*in C eine sehr gute Leistung aufweist, da sie in jedem Test jeweils mehr Punkte erzielt hat als alle ihre Mitschüler*innen. Zum anderen wird Schüler*in B eine sehr gute Leistung attestiert, da sie den größten Leistungszuwachs zeigt. Schließlich ist auch noch zu beobachten, dass das Erzielen von z. B. mindestens 10 korrekten Sätzen als eine gute Leistung bewertet wird, da dies z. B. einem im Bildungsplan verankerten Ziel entspräche.

2.2. Kriteriale, individuelle und soziale Bezugsnorm

Diese drei Argumentationsstränge werden in den Bildungswissenschaften Bezugsnormen (engl. „reference norms“) genannt. Darunter versteht man den „Standard“, der zur Bewertung einer Leistung herangezogen wird (Heckhausen, 1974). Die „soziale Bezugsnorm“ liegt vor, wenn das zu bewertende Ergebnis mit den Ergebnissen einer Vergleichsgruppe, die denselben Test durchlaufen hat, abgeglichen wird. Die „kriteriale Bezugnorm“ (auch „sachliche Bezugnorm“; Rheinberg & Fries, 2010) legt a priori fest, was eine „gute“ Leistung auszeichnet. Ein typisches Beispiel aus dem Schulalltag sind etwa Wertungstabellen des Sportunterrichts. Dort ist bspw. festgelegt, dass eine Oberstufenschülerin 14 Notenpunkte für 200m Rückenschwimmen bekommt, wenn sie weniger als 223,5 Sekunden aber mehr als 217,7 Sekunden dafür benötigt (Ministerium für Kultus, Jugend und Sport Baden-Württemberg, 2022). Schließlich wird in der Literatur noch die „individuelle Bezugsnorm“ (auch „temporale“, „intraindividuelle“ oder „ipsative Bezugsnorm“) genannt. Sie vergleicht (wie die soziale Bezugsnorm) zur Bewertung einer Leistung diese ebenfalls mit anderen realen (empirisch gemessenen) Leistungen. Allerdings stammen jene Leistungen vom selben Merkmalsträger (Individuum), sie wurden aber zu einem früheren Zeitpunkt erfasst (siehe Abbildung 1).

Abbildung 1: Veranschaulichung der Bezugsnormen. Die Datengrundlage entspricht Tabelle 1.

2.3. Die Relativität und Normativität der Bezugsnormen

Die drei Bezugsnormen unterscheiden sich – wie oben angeführt – im Standard, auf den sie sich zur Bewertung einer einzelnen Leistung beziehen. Damit scheint der Wahl der Bezugsnorm ein stark normatives Moment innezuwohnen, definiert sie doch einen jeweils unterschiedlichen Idealzustand. Jedoch erfolgt in allen Bezugsnormen die Bewertung anhand von Abständen auf der Leistungsvariablen (siehe Abbildung 1). Im Folgenden wird daher dafür argumentiert, dass alle drei Bezugsnormen Relativierungen darstellen, die sich von der sozialen Bezugsnorm lediglich um eine Translation (kriteriale Bezugsnorm) oder bezüglich der zugrunde gelegten Variablen unterscheiden (individuelle Bezugsnorm). Dies soll am Beispiel aus Tabelle 1 für alle drei Bezugsnormen illustriert werden.

Legt sich die Grundschullehrkraft aus dem einführenden Beispiel normativ auf die individuelle Bezugsnorm fest, ergibt die Relativierung des letzten Messwertes von Schüler*in C am ersten Messwert eine Steigerung um 3 Sätze, wohingegen sich Schüler*in F um 4 Sätze gesteigert hat. Die Lehrkraft muss nun nicht nur normativ entscheiden, welche dieser Leistungssteigerungen einer „wie guten" Leistung entspricht. Es stellt sich darüber hinaus die Frage, wie differenziert diese Transformation sein soll. Macht das Ausgangsniveau der Leistung einen Unterschied? Oder ob Deutsch die Zweit- oder Erstsprache der Schüler*in ist? Es ist nicht plausibel anzunehmen (wenngleich nach unserem Wissen nicht empirisch belegt), dass Lehrende diese Fragen via a priori abgeleiteter Prinzipien beantworten. Vielmehr ist zu erwarten, dass bei dieser Bewertung die Leistungssteigerung eines Individuums am Erfahrungshorizont von bisher wahrgenommenen Leistungssteigerungen relativiert wird. Danach hätte also eine Schüler*in unter Annahme der individuellen Bezugsnorm eine „gute Leistung" erbracht, wenn ihre Leistungssteigerung *im Vergleich zu Leistungssteigerungen implizit assoziierter anderer* überdurchschnittlich ist. Damit erfolgt auch unter Nutzung der individuellen Bezugsnorm die Leistungsbewertung durch soziale Relativierung.

Die Nutzung einer kriterialen Bezugsnorm scheint zunächst frei von einer solchen interindividuellen Relativierung, da es sich ja um eine Idealnorm handelt (Klauer et al., 1972) und bei dieser die Transformation des Messwerts zur Bewertung (Abstand Messwert vs. Kriterium; siehe Pfeil 3 in Abbildung 1) „in der Sache selbst liegt" (Rheinberg & Fries, 2010, S. 61). Dass dies jedoch praktisch kaum der Fall sein kann, wird sehr schnell plausibel: Angenommen, eine Lehrkraft weiß qualitativ alles über den Lerngegenstand – also in unserem Beispiel über den Schriftspracherwerb (z.B. diverse Stufenmodelle, effektive Unterrichtspraktiken, typische Lernschwierigkeiten, etc.) – aber nichts über typische quantitative Resultate: Sie würde kaum ableiten und begründen können, warum welche Anzahl an richtigen Sätzen als „gute Leistung" bewertet werden sollten. Vielmehr ist plausibel, dass bei der Festlegung der Schwellenwerte Wissen über die Resultate (wer macht mit welchen Ergebnissen typischerweise guten Fortschritt in der nächsten Klassenstufe etc.) eine zentrale Rolle spielen. Damit definiert wiederum eine soziale Relativierung, was eine „gute Leistung" ausmacht. Im Unterschied zur sozialen Bezugsnorm stammen die Resultate, anhand derer die Relativierung stattfindet, jedoch auch von Merkmalsträgern außerhalb der zu bewertenden Schüler*innen. Einschränkend muss an dieser Stelle jedoch ergänzt werden, dass nach unserem Wissen kaum tragfähige empirische Untersuchungen bei Lehrpersonen vorliegen, die der Frage der sozialen Relativierung bei der Setzung von Referenzwerten für kriteriale Bezugsnormen nachgehen und die hier aufgestellten Hypothesen verifizieren oder falsifizieren könnten.

3. Relativität und Normativität in der unterrichtspraktischen Leistungsbewertung

In der unterrichtlichen Praxis der (summativen) Leistungsbewertung findet man kaum die ausschließliche Anwendung genau einer Bezugsnorm. Vielmehr herrscht Einigkeit in der Literatur, dass in der unterrichtlichen Praxis die Anwendung mehrerer Bezugsnormen bei der Messung und Bewertung von Leistungen weit verbreitet ist (Bohl, 2004; Lintorf & Buch, 2021; Pant, 2020), was sich erstaunlicherweise mit vielen Abschnitten aus Verordnungen und Gesetzen der Bundesländer zur Notenvergabe (Pant, 2020) deckt.

Wie schnell eine Vermischung von Bezugsnormen in der unterrichtlichen Praxis erfolgen kann, soll folgendes Beispiel illustrieren: Eine Lehrkraft unterrichtet eine Einheit und schließt diese mit einer summativen Leistungserfassung (z.B. einer Klassenarbeit) ab. Die Unterrichtseinheit deckt sich im Großen und Ganzen mit dem Bildungsplan, lediglich ein einzelner im Bildungsplan geforderter Teilaspekt wird im Unterrichtsverlauf nur gestreift. Der Definition der Noten gemäß orientiert sich die Lehrkraft bei der Erstellung der Klassenarbeit am Bildungsplan und wählt je Teilaspekt im Bildungsplan zwei Aufgaben für die Klassenarbeit aus (kriteriale Bezugsnorm). Da die so konstruierte Klassenarbeit aus Sicht der Lehrkraft zu viele Aufgaben enthält, streicht sie diejenigen beiden Aufgaben, deren Inhalte im Unterricht zuvor nur sehr kurz behandelt wurden (Verschiebung des Kriteriums). Danach wählt die Lehrkraft die maximal zu erreichenden Punkte für jede Aufgabe, wobei einfache Aufgaben (im Sinne hoher erwarteter Lösungshäufigkeiten) weniger Punkte erhalten (soziale Bezugsnorm). Abschließend ergänzt die Lehrkraft die Klassenarbeit um eine Zusatzaufgabe (Abweichung von der kriterialen Norm). Diese antizipiert die Lehrkraft als eher schwierig, versieht sie aber dennoch mit wenigen Punkten (Abweichung von der sozialen Norm). Nach der Korrektur legt die Lehrkraft zunächst ihren üblichen „Notenschlüssel“ an und berechnet den Durchschnitt. Dieser erscheint ihr in Anbetracht der wahrgenommenen Lernfortschritte während der Unterrichtseinheit „zu schlecht“, weshalb sie den Schlüssel abändert (individuelle Bezugsnorm).

4. Relativität und Normativität in der Bildungsforschung

Auch in Verfahren der Leistungsmessung, die in der empirischen Bildungsforschung zum Einsatz kommen, spielen Normativität und Relativität eine erhebliche Rolle. Da an dieser Stelle keine Einführung in die Psychometrie gegeben werden kann, werden stattdessen einige konzeptuelle Aspekte anhand eines Beispiels illustriert und auf vertiefende Literatur verwiesen.

Ein erstes normatives Moment der Leistungsmessung in der Bildungsforschung wird bereits bei der Wahl der metatheoretischen Grundannahmen sichtbar: Die überwiegende Mehrheit der empirisch-quantitativen Arbeiten nimmt an, mit der „Leistung“ eine latente Variable – also eine Variable, die nicht direkt, sondern nur mittelbar über Indikatoren gemessen werden kann (Borsboom, 2008) – zu messen (Edelsbrunner & Dablander, 2019). Dazu wird ein mathematisches Modell spezifiziert, das den Zusammenhang zwischen der empirischen Datenstruktur und der

latenten Variablen („Leistung“) beschreibt. Diese Modellierung erfolgt derzeit für fachliche Kompetenzen meist anhand unidimensionaler Item Response Modelle (IRT; Kelava & Moosbrugger, 2020). Diese postulieren unter anderem, dass die Wahrscheinlichkeit, dass eine Schüler*in p eine Testfrage i richtig beantwortet $P(X_{pi} = 1)$, lediglich von der Schwierigkeit der Frage β_i und der Ausprägung der latenten Variable (Leistung) der Schüler*in θ_p abhängt, wobei diese Größen wie folgt in Beziehung stehen.

$$P(X_{pi} = 1) = \frac{e^{\theta_p - \beta_i}}{1 + e^{\theta_p - \beta_i}}$$

Andere Variablen (z.B. weitere Merkmale der Schüler*innen) oder Parameter, die die $P(X_{pi} = 1)$ beeinflussen, werden als nicht-existent angenommen. Diese Annahme des Modells, welche auch Rasch-Homogenität genannt wird, trägt insofern Normativität in den Prozess der Leistungsmessung ein, als dass sie mitbestimmt, welche Aufgaben der Leistungstest final enthält: Denn Aufgaben, deren empirische Datenmuster der Annahme der Raschhomogenität widersprechen, werden im Prozess der Testkonstruktion entfernt bzw. ersetzt. Die Verwendung von IRT gilt im Kontext von Large Scale Assessments als wissenschaftlicher Konsens (Borsboom & Wijsen, 2017). Nichtsdestotrotz stellt diese Wahl selbst bereits eine normative Entscheidung dar, da sie wie beschrieben mit bestimmten Annahmen einhergeht. Genauso würde die Entscheidung für eine alternative Modellierung (z.B. Knowledge Space Theory [KST]; Falmagne et al., 1990) eine normative Entscheidung darstellen, da diese ebenfalls (wenngleich auch andere) Annahmen trifft.

Relativität wird bei der Messung von Leistung anhand IRT dadurch erzeugt, dass die Fähigkeitsparameter θ_p einer Normalverteilung folgen (müssen). Damit kann aus Kenntnis des Erwartungswertes der θ_p sowie deren Streuung jedes θ_p in Perzentile transformiert werden (siehe Abbildung 2). In Abbildung 2 wird zudem deutlich, dass eine Schüler*in X, die bei diesem Leistungstest 425 Punkte erzielt hat, 26% aller Schüler*innen übertrifft. Obwohl die IRT damit (im Gegensatz zur KST) die Anwendung einer sozialen Bezugsnorm impliziert, erlaubt ein solcher Konstruktionsprozess der Kompetenzmodellierung auch die Anwendung der individuellen und kriterialen Bezugsnorm: Die individuelle Bezugsnorm etwa wird besonders einfach anwendbar, da die Leistung einer Person mit nur einem Parameter θ_p beschrieben werden kann. Durch die sog. Verankerung von Items wird es sogar möglich, θ_p anhand überschneidungsfreier Aufgabensets zu bestimmen und damit Lernfortschritte in der Einheit von θ_p zu bestimmen. Die Anwendung der kriterialen Bezugsnorm wird durch die Bestimmung sog. Ankerpunkte (Beaton & Allen, 1992) für die β_i ermöglicht. Die Bestimmung dieser Ankerpunkte stellt eine Post-Hoc Analyse (Analyse nach Erhebung der Daten) von Iteminhalten dar, in der fachdidaktische Expert*innen das Leistungskontinuum in diskrete Intervalle unterteilen (Kompetenzstufen, siehe Abbildung 2). Dabei werden Aufgabeninhalte und Aufgabenanforderungen derjenigen Aufgaben untersucht, die

empirisch gesehen besonders spezifisch für einen Ankerpunkt auf der Kompetenzskala bzgl. notwendiger kognitiver Operationen, inhaltlicher Kriterien (z.B. Wortschatz eines Lesetextes), Aufgabenformate etc. sind (Hartig, 2007). Die Schnittmenge dieser Aufgabencharakteristika wird dann zur Definition der Kompetenzstufe herangezogen. Da die Schwierigkeitsparameter β_i und die Fähigkeitsparameter θ_p dieselbe Metrik aufweisen („auf derselben Achse liegen"), kann die Leistung θ_p kriterial im Sinne von „ist fähig, Aufgaben der Art/Kompetenzstufe X zu lösen" interpretiert werden.

Abbildung 2: Relativität und Normativität in der in der Kompetenzstufenmodellierung

5. Fazit

Eine Leistungsbewertung erfordert notwendigerweise die Relativierung eines Messwertes. Im Beitrag wurde argumentiert, dass Lehrkräfte diese Relativierung (implizit) entlang normativ gewählten Bezugsnormen vornehmen, wobei diese im Prozess oftmals wechselt, während in psychometrischen Modellierungen von schulischer Leistung bereits die Wahl des mathematischen Modells eine normative (wenngleich auch gut begründete) Entscheidung darstellt. Relativierungen von Leistungen sind in den häufig vorkommenden Modellen der IRT sowohl sozial als auch kriterial und intraindividuell leicht möglich und inhaltlich interpretierbar.

6. Literatur

Beaton, A. E., & Allen, N. L. (1992). Interpreting Scales Through Scale Anchoring. Journal of Educational Statistics, 17(2), 191. https://doi.org/10.2307/1165169

Bohl, T. (2004). Prüfen und Bewerten im Offenen Unterricht. Beltz.

Borsboom, D. (2008). Latent variable theory. Measurement: Interdisciplinary Research & Perspective, 6(1–2), 25–53. https://doi.org/10.1080/15366360802035497

Borsboom, D., & Wijsen, L. D. (2017). Psychology's atomic bomb. Assessment in Education: Principles, Policy & Practice, 24(3), 440–446. https://doi.org/10.1080/0969594X.2017.1333084

Edelsbrunner, P. A., & Dablander, F. (2019). The Psychometric Modeling of Scientific Reasoning: A Review and Recommendations for Future Avenues. Educational Psychology Review, 31(1), 1–34. https://doi.org/10.1007/s10648-018-9455-5

Falmagne, J.-C., Doignon, J.-P., Koppen, M., Villano, M., & Johannesen, L. (1990). Introduction to knowledge spaces: How to build, test, and search them. Psychological Review, 97(2), 201–224. https://doi.org/10.1037/0033-295X.97.2.201

Hartig, J. (2007). Skalierung und Definition von Sprachniveaus. In B. Beck & E. Klieme (Hrsg.), Sprachliche Kompetenzen: Konzepte und Messung. Beltz.

Heckhausen, H. (1974). Leistung und Chancengleichheit. Hogrefe.

Holmeier, M. (2013). Leistungsbeurteilung im Zentralabitur. Springer VS.

Klauer, K. J., Fricke, R., Herbig, M., Rupprecht, H., & Schott, F. (Hrsg.). (1972). Lehrzielorientierte Tests. Schwann.

Lintorf, K., & Buch, S. R. (2021). Stabile Präferenz oder flexibel am Diagnoseziel orientiert? – Die Bezugsnormwahl angehender Lehrkräfte. Zeitschrift für Pädagogische Psychologie, 35(2–3), 107–118. https://doi.org/10.1024/1010-0652/a000271

Kelava, A., & Moosbrugger, H. (2020). Einführung in die Item-Response-Theorie (IRT). In H. Moosbrugger & A. Kelava (Hrsg.), Testtheorie und Fragebogenkonstruktion (S. 369–409). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-61532-4_16

Pant, H. A. (2020). Notengebung, Leistungsprinzip und Bildungsgerechtigkeit. In S.-I. Beutel & H. A. Pant (Hrsg.), Lernen ohne Noten. Alternative Konzepte der Leistungsbeurteilung. Kohlhammer.

Rheinberg, F. (1980). Leistungsbewertung und Lernmotivation. Hogrefe.

Rheinberg, F., & Fries, S. (2010). Bezugsnormorientierung. In D. H. Rost (Hrsg.), Handwörterbuch pädagogische Psychologie (4. Aufl.). Beltz.

Rost, J. (2004). Lehrbuch Testtheorie - Testkonstruktion. Huber.

Wilbert, J., & Gerdes, H. (2009). Die Bezugsnormwahl bei der Bewertung schulischer Leistungen durch angehende Lehrkräfte des Förderschwerpunktes Lernen. Heilpädagogische Forschung, 35, 122–135.

Angaben zu den Autoren:

- Prof. Dr. Samuel Merk, Juniorprofessor für Empirische Schul und Unterrichtsforschung, Pädagogische Hochschule Karlsruhe, samuel.merk@ph-karlsruhe.de

- Sarah Bez, wissenschaftliche Mitarbeiterin, Institut für Erziehungswissenschaft an der Universität Tübingen und Institut für Schul- und Unterrichtsentwicklung an der Pädagogischen Hochschule Karlsruhe, sarah.bez@uni-tuebingen.de