

Beurteilung von Bewertungskompetenz: Entwicklung und Evaluation von Beurteilungsmodellen

Einleitung

Bewertungskompetenz und Argumentieren im Physikunterricht (PU) zu fördern, scheint auch fünfzehn Jahre nach Einführung des Kompetenzmodells in Österreich vielen Lehrkräften schwer zu fallen. Das führt dazu, dass die Vermittlung dieser Kompetenz im PU oftmals zu kurz kommt. Auch bei Leistungsfeststellungen jeglicher Art sind Aufgabenstellungen dazu unterrepräsentiert. Gründe dafür werden darin verortet, dass einerseits der Pool an bestehenden Aufgaben klein ist. Andererseits ist bei Lehrkräften Verunsicherung zu bemerken, wie Bewertungskompetenz in einer transparenten und objektiven Form beurteilt werden kann, da es hierzu kaum Modelle gibt, die für die tägliche Anwendung in der Schule viabel sind.

Der vorliegende Beitrag greift Design-based Research Studien auf (Ejersbo et al., 2008) und zeigt Anforderungen an mögliche Aufgabenstellungen für die Sekundarstufe I und II, die parallel und passend zu Beurteilungsmodellen (5-stufig und 3-stufig) entwickelt wurden und inhaltliche und sprachliche Aspekt getrennt bewerten. Darüber hinaus werden erste Ergebnisse präsentiert, inwieweit die Beurteilung und Feedback zu Schülerarbeiten von einer KI am Beispiel ChatGPT übernommen werden kann.

Bewertungskompetenz

Bewertungskompetenz im naturwissenschaftlichen Unterricht bezeichnet die Fähigkeit, fachliche Sachverhalte und Handlungsoptionen mithilfe von Kriterien aus fachlicher, ökologischer, ökonomischer, gesellschaftlicher und ethischer Perspektive zu prüfen, Evidenzen zu gewichten und auf dieser Grundlage begründete Urteile oder Entscheidungen zu formulieren (KMK, 2020). Diese Kriterien lassen sich aus den drei Säulen der Bildung nachhaltiger Entwicklung ableiten (Strange & Bayley, 2008). Das umfasst nach dem Göttinger Modell Teilkompetenzen wie das Erkennen von Bewertungsanlässen, das Abwägen konkurrierender Argumente sowie die Reflexion der eigenen Urteilsfindung (Eggert & Bögeholz, 2006). Das wird in der Physikdidaktik als Kriterium zwischen rational-analytischer und intuitiver Entscheidungsweise diskutiert (Sander & Höttecke, 2015). Im österreichischen Lehrplan des Gymnasiums (Allgemein bildende höhere Schule) ist dazu ausgeführt, dass Lernende „[...] die Fähigkeit erwerben sollen, Fähigkeit, naturwissenschaftlich begründet zu argumentieren und am gesellschaftlichen Diskurs teilzunehmen“ (BMB, 2025). Operationalisiert wird das durch vier Punkte, von denen hier zwei herausgegriffen werden: Lernende sollen demnach „Bedeutung, Chancen und Risiken der Anwendungen von naturwissenschaftlichen Erkenntnissen auf persönlicher, regionaler und globaler Ebene erkennen, um verantwortungsbewusst zu handeln [...] und] Entscheidungskriterien für das eigene Handeln entwickeln und aus naturwissenschaftlicher Sicht überprüfen“ (ebd.). Dabei sollen sie lernen, „[...] Informationen aus unterschiedlich verlässlichen Quellen aus naturwissenschaftlicher Sicht und aus anderen Blickwinkeln (z.B. ökonomisch, ökologisch, ethisch) [zu] reflektieren“ (ebd.).

Das 5-stufige Modell: Entwicklung und Erprobung

Will man nun Lernende dazu anleiten, die oben beschriebenen Kompetenzen zu erwerben, muss der Unterricht in den Schulen entsprechend gestaltet sein. Dazu bedarf es zweier Dinge: Einerseits geeigneter Aufgaben, da diese ein wesentliches Steuerelement des Unterrichts sind (Bezler, 2002; Kauertz et al., 2015). Andererseits benötigen Lehrpersonen einen Rahmen innerhalb dessen sie imstande sind, diese Aufgaben zu bewerten und den Lernenden Feedback dazu zu geben. Hinsichtlich geeigneter Aufgaben stellte sich in einer Literaturrecherche heraus, dass erstaunlich wenige geeignete zu finden sind (Gstall, 2024). Die normativ festgesetzten Anforderungen an geeignete Aufgaben beinhalten, dass die möglichst wenig Lesekompetenz abfragen sollen, zumindest was die Sek I betrifft. Weiters soll nicht das Fachwissen abgeprüft werden, sondern der Fokus soll auf dem Argumentieren, also auf der eigentlichen Bewertungskompetenz liegen. Darüber hinaus liegen für den praktischen Gebrauch in der Schule kaum geeignete Beurteilungsmodelle vor. So erscheint zum Beispiel das Göttinger Modell (Eggert & Bögholz, 2006) als zu komplex für den Schulalltag, während es, wie auch das Modell der Scientific Literacy (Prenzel et al., 2001), sprachliche Kriterien unberücksichtigt lässt. Aus beiden Modellen lässt sich aber übernehmen, dass aufeinander aufbauende Niveaustufen sinnvoll sind. Das Modell nach Hostenbach et al. (Hostenbach et al., 2011) lässt sich wiederum inhaltlich und strukturell mit dem oben beschriebenen Kompetenzmodell der Naturwissenschaften bestens verbinden und berücksichtigt sprachliche Aspekte. Gstall (2024) entwickelte auf Basis u.a. dieser Modelle ein fünfstufiges Modell, das beides, inhaltliche und sprachlich-syntaktische Aspekte, abbildet.

Passend dazu wurden vier Aufgabenstellungen entwickelt, die Freitextantworten verlangen, um die Argumentationsfähigkeit abzufragen. Der Fokus lag hier auf der Sek I. Um die Textmenge zu reduzieren, wurden Informationen in Tabellenform vorgegeben. In jeder Tabelle fanden sich mindestens drei Kriterien, anhand derer entschieden werden sollte, z.B. welches Elektrogerät angeschafft werden soll.

In einer Studie mit $N_{\text{SekI}} = 81$ Proband:innen der 6. und 7. Schulstufe wurden drei Aufgabenstellungen erprobt und anschließend $N_L = 19$ Lehrpersonen und Fachdidaktiker zur Beurteilung mit dem Raster übergeben. Mittels offener Fragebogen wurde dessen Viabilität erfragt. Hauptaussagen waren der Wunsch nach einem noch einfacheren Modell sowie der Möglichkeit, inhaltliche und sprachliche Aspekte getrennt zu beurteilen (Gstall, 2024).

Aufgaben für die Sek II, 3-stufiges Modell und Beurteilung mit KI

Auf Basis dieser Rückmeldungen wurde ein 3-stufiges Modell entwickelt (Abb. 1), das auch insofern adaptiert wurde, dass es auch für Sek II passend sein soll (Hackner, 2025). Hinsichtlich der Aufgabenstellungen lag der Schwerpunkt auf der Weiterentwicklung anspruchsvollerer Formate für die Sek II. Dabei sollten die Informationen zu den zu beurteilenden Kriterien nicht so offensichtlich in Tabellenform angegeben sein, sondern ähnlicher zu lebensweltlichen Situationen gestaltet, in kurzen Informationstexten gegeben sein. Teil a) der Aufgaben verlangt nach einer Extraktion der Informationen aus den Texten, während in Teil b) die eigentliche Bewertung stattfinden soll. Die Vorgehensweise wurde so gewählt, um die unterschiedlichen Kompetenzbereiche unterschiedlichen Teile der Aufgaben zuordnen zu können und um Fachwissen, bzw. Lesekompetenz klar von Argumentationsfähigkeit zu trennen. Es fand ein Durchlauf mit 6 Aufgabenstellungen statt, die von $N_{\text{SekII}} = 125$ Lernenden der 11. Schulstufe bearbeitet wurden und anschließend 5 Lehrkräften zur Beurteilung gegeben wurden. Ergebnisse stehen derzeit noch aus, allerdings

Niveau	Anzahl der Kriterien	Inhaltliche Anforderungen	Anker-Beispiele
0/-	0 - 1	Die Begründung(en) (falls vorhanden) wurden auf Basis naturwissenschaftlichen Alltagswissens aufgebaut und sind deswegen fachlich unangemessen. Die Entscheidung wurde intuitiv getroffen. Einiges an Interpretationsspielraum und/oder die gezogene Schlussfolgerung ist falsch bzw. nicht zielführend oder für die Beantwortung der Aufgabenstellung irrelevant und/oder möglicherweise wurden eigene, irrelevante Kriterien in die Argumentation miteinbezogen.	„Ich wähle Option A, sie ist ziemlich günstig.“ „Ich wähle Option A, weil sie ziemlich günstig ist und mir die Farbe gefällt.“ (1 Kriterium, z.B. Farbe eines Objekts ist irrelevantes Kriterium)
1/-	≥ 2	Kriterien, die in der Aufgabenstellung als wesentlich vermerkt wurden (z.B. geringer Preis: Sara möchte beim Kauf nicht viel Geld ausgeben), werden in die Argumentation miteinbezogen. aber es kann noch Interpretationsspielraum geben, dass noch mehr als die explizit genannten Kriterien Einfluss auf die Wahl der Schüler*innen hatten. und/oder möglicherweise wurden eigene Kriterien in die Argumentation miteinbezogen, die irrelevant sind.	„Ich wähle Option A, da sie am günstigsten ist und die höchste Leistung hat.“ „Ich wähle Option A, da sie am günstigsten ist und die höchste Leistung hat. Außerdem ist sie praktisch.“ (2 Kriterien + 1 angedeutet)
2/+	≥ 2	Die Argumentation für eine Option ist durchgehend verständlich. Es gibt keinen Interpretationsspielraum oder Zweifel. und Die vorliegenden Kriterien und Optionen werden genutzt, um auch konträre Standpunkte zu entkräften. und/oder Mehrere Kriterien und/oder Optionen werden im Laufe der Argumentation berücksichtigt und begründet gegeneinander aufgewogen. Kriterien, die beim Vergleichen verwendet werden, zählen auch zur Anzahl der Kriterien.	„Ich wähle Option A, da sie im Vergleich zu den anderen am günstigsten ist und auch viele Funktionen bietet. Zwar bietet Option B eine höhere Leistung, der Unterschied ist jedoch gering, weshalb auch Option A zufriedenstellende Ergebnisse liefern wird.“ (3 Kriterien)
Niveau	Sprachliche Merkmale		Anker-Beispiele
0/-	Die Begründung kann sprachlich erkennbar sein. Die Argumentation folgt jedoch meist einer einfachen Struktur wie „Option A ist die beste, weil Kriterium 1.“		„Ich wähle Option A, sie ist ziemlich günstig.“ „Ich wähle Option A, weil sie ziemlich günstig ist und mir die Farbe gefällt.“
1/-	Die Argumentation geht über folgende Struktur hinaus: „Option A ist die beste, weil Kriterium 1 + 2.“ Und beinhaltet gegebenenfalls weitere Kriterien und/oder Adverbien wie außerdem/außer/zusätzlich/.../		„Ich wähle Option A, da sie am günstigsten ist und die höchste Leistung hat.“ „Ich wähle Option A, da sie am günstigsten ist und die höchste Leistung hat. Außerdem ist sie praktisch.“
2/+	In der Argumentation ist klar erkennbar, dass Optionen sprachlich miteinander verglichen werden, zum Beispiel durch Ausdrücke wie zwar / aber / jedoch / trotzdem / ...		„Ich wähle Option A, da sie im Vergleich zu den anderen am günstigsten ist und auch viele Funktionen bietet. Zwar bietet Option B eine höhere Leistung, der Unterschied ist jedoch gering, weshalb auch Option A zufriedenstellende Ergebnisse liefern wird.“

Abbildung 1: 3-stufiges Beurteilungsmodell, Oberstufe, inhaltliche und sprachliche Anforderungen getrennt (Hackner, 2025)

fiel auf, dass teilweise die Informationen von den Lernenden gut extrahiert wurden, die Argumentation trotzdem ohne jedes Faktum erfolgte.

Parallel dazu fand eine weitere Studie statt, die die Antworten der Lernenden mit einer KI (Chat-GPT 4o) nach beiden oben erwähnten Rastern beurteilen ließ (Humer, 2025). Dabei lag der Fokus auf zwei Aspekten: Erstens, Prompts nach dem P.R.O.P.E.R Framework (Gruber, 2024) zu generieren, die zufriedenstellende, reliable Ergebnisse liefern. Andererseits wurde die KI genutzt, um Feedback für die Lernenden generiert, dessen Qualität untersucht wurde. Ein vorläufiges Ergebnis ist, dass die Beurteilungen von Chat-GPT nur teilweise reproduzierbar sind, aber vergleichbar mit der Spanne der Beurteilungen durch Lehrpersonen. Mensch wie Maschine zeigen vor allem im mittleren Leistungsspektrum stärkere Schwankungen als bei sehr guten oder sehr schlechten Arbeiten. Hier ergibt sich allerdings die Möglichkeit, die Beurteilung durch die KI mehrmals wiederholen zu lassen und Mittelwerte zu bilden.

Ausblick

Die drei berichteten Studien legen nahe, dass Argumentieren im naturwissenschaftlichen Unterricht geübt werden muss, um hier Fortschritte zu generieren. Speziell wünschenswert erscheint es, Lernende zu einer faktenbasierten Argumentation zu erziehen. Dazu bedarf es offengelegter Kriterien, was in dieser Arbeit mit den entwickelten Beurteilungsrastern erfüllt wird. Diese sollen sowohl für die Lernenden als auch für die Lehrenden Sicherheit bieten. Allerdings befinden wir uns im Spannungsfeld zwischen genauen Beurteilungskriterien, die einschränkend wahrgenommen werden können, und möglichst vielfältigen, offenen Aufgaben. Es ist angedacht, das Beurteilungsraster so weiterzuentwickeln, dass in Zukunft auch Aufgaben beurteilt werden können, bei denen auch die Informationen selbst recherchiert werden müssen.

Literatur

- Bezler, H. J. (2002). *Intelligentes Üben - Behalten - Überprüfen. Positionen einer neuen „Aufgabenkultur“ im naturwissenschaftlichen Unterricht. Materialien zur Bildungsforschung Qualitätsentwicklung im naturwissenschaftlichen Unterricht*. Fachtagung am 15. Dezember 1999, Frankfurt/Main.
- BMB. (2025). *Lehrpläne – allgemeinbildende höhere Schulen*. BMB. Quelle: <https://www.ris.bka.gv.at/NormDokument.wxe?Abfrage=Bundesnormen&Gesetzesnummer=10008568&Artikel=&Paragraf=&Anlage=1&Uebergangsrecht=> [Zugriff: 2015-10-27]
- Eggert, S. & Bögholz, S. (2006). Göttinger Modell der Bewertungskompetenz – Teilkompetenz „Bewerten, Entscheiden und Reflektieren“ für Gestaltungsaufgaben Nachhaltiger Entwicklung. *ZfdN*, 12, 177-197.
- Ejersbo, L. R., Engelhardt, R., Frølund, L., Hanghøj, T., Magnussen, R. & Misfeldt, M. (2008). Balancing product design and theoretical insights. In A. E. Kelly, R. A. Lesh, & J. Y. Baek (Eds.), *The Handbook of Design Research Methods Education* (pp. 149-163). Routledge.
- Gruber, P. H. (2024). *Using ChatGPT for Advanced Data Analysis. Lecture notes*. Università della Svizzera italiana.
- Gstall, D. (2024). *Wie bewertet man Bewertungskompetenz? Entwicklung und Evaluation eines Beurteilungsmodells zur S-Kompetenz für die Unterstufe* [Masterarbeit]. Universität Wien. Wien.
- Hackner, K. (2025). *Bewertung der S-Kompetenz. Überarbeitung eines Modells* [Masterarbeit]. Universität Wien. Wien.
- Hostenbach, J., Fischer, H. E., Kauertz, A., Mayer, J., Sumfleth, E. & Walpuski, M. (2011). Modellierung der Bewertungskompetenz in den Naturwissenschaften zur Evaluation der Nationalen Bildungsstandards. *ZfdN*, 17, 261-288.
- Humer, M. (2025). *Die Bewertung von Bewertungskompetenz mithilfe des Sprachmodells ChatGPT* [Masterarbeit]. Universität Wien. Wien
- IQS. (2021). *IKM – Physik, Beginn 8. SSt. Überblick Aufgabenpaket mit Ansichtsexemplar und Detailinformationen*. Quelle: <https://www.iqs.gv.at/themen/nationale-kompetenzerhebung/online-plattform-ikm-plus> [Zugriff: 2024-11-03]
- Kauertz, A., Löffler, P. & Fischer, H. E. (2015). Physikaufgaben. In E. Kirchner, R. Girwidz, & P. Häußler (Eds.), *Physikdidaktik. Theorie und Praxis* (pp. 451-478). Springer-Verlag.
- KMK. (2020). *Bildungsstandards im Fach Physik für die Allgemeine Hochschulreife*. I. z. Q. i. Bildungswesen. https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2020/2020_06_18-BildungsstandardsAHR_Biologie.pdf [Zugriff: 2025-10-27]
- Prenzel, M., Rost, J., Senkbeil, M., Häußler, P. & Klopp, A. (2001). Naturwissenschaftliche Grundbildung: Testkonzeption und Ergebnisse. In D. PISA-Konsortium (Ed.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (pp. 191 - 248). Leske + Budrich.
- Sander, H. & Höttecke, D. (2015). Bewertungskompetenz in der Physikdidaktik: Zwischen Rationalität und Intuition. In F. H., S. H., & W. H. (Eds.), *Fachlich argumentieren lernen* (pp. 153 - 168). Waxmann.
- Strange, T. & Bayley, A. (2008). *Sustainable Development. Linking economy, society, environment*. OECD. Quelle: https://www.oecd.org/content/dam/oecd/en/publications/reports/2008/12/sustainable-development_g1gh9be9/9789264055742-en.pdf [Zugriff: 2025-10-27]