Jannis Zeller[1] [1]Universität Paderborn
Josef Riese[1]

Fähigkeitsprofile im Physikdidaktischen Wissen mithilfe von Machine Learning

Theoretischer Hintergrund und Motivation

Informatives formatives Feedback besitzt das Potenzial, Lernprozesse anzustoßen und positiv zu beeinflussen (z. B. Hattie & Timperley, 2007; Shute, 2008). Obwohl das Professionswissens angehender Lehrkräfte bereits seit einigen Jahren im Fokus fachdidaktischer Forschung steht (Kaiser, Bremerich-Vos & König, 2020), gibt es bisher im deutschsprachigen Raum jedoch kaum Ansätze, diese Forschungsergebnisse auch in nützliche Feedback- bzw. Assessment-Instrumente umzusetzen. Existierende Konzepte (z. B. Jordans et al., 2022) nutzen meist Multiple-Choice-Testinstrumente, deren Validität im Vergleich zu offeneren Testinstrumenten kritisch eingeschätzt werden kann. Gerade für die Professionswissensdomäne des fachdidaktischen Wissens (FDW), die als „amalgam" (Shulman, 1986) ein stark vernetztes und häufig weniger explizierbares Konstrukt darstellt (siehe auch „personal Pedagogical Content Knowledge", z. B. Alonzo et al., 2019), wäre ein Feedback-Tool auf Basis umfangreich validierter offener Testinstrumente (z. B. Gramzow, 2015; Kröger, 2019) wünschenswert.

Um inhaltlich informatives Feedback, d. h. Feedback, welches über die reine Angabe eines Scores hinausgeht, zu ermöglichen, sind kriterienorientierte Beschreibungen des Zielkonstrukts notwendig. Solche Beschreibungen von Ausprägungen des FDW angehender Physiklehrkräften existieren bisher primär in Form von Scale-Anchoring-Modellierungen (Schiering et al., 2023; Zeller et al., 2022). Diese auf Item-Response-Modellierungen basierenden Niveaumodelle erlauben allerdings primär stark generalisierte, strikt hierarchische Aussagen über Wissensausprägungen. Sie lieferten Hinweise, dass sich das FDW in niedrigen Ausprägungen auf reproduktive Aspekte beschränkt und sich in höheren Ausprägungen hin zu evaluierenden und kreativen Elementen erweitert.

In diesem Beitrag soll untersucht werden, ob sich mithilfe von Machine Learning (ML) und Natural Language Processing (NLP) auch weitere, vor allem nicht hierarchische prototypische Ausprägungen („Fähigkeitsprofile") des FDW in einem großen Datensatz zum FDW aus dem Projekt ProfiLe-P+ (Vogelsang et al., 2019) finden lassen. Dazu werden im Sinne einer Computational Grounded Theory (Nelson, 2020) authentische Sprachproduktionen der Proband:innen mit menschlichem Expertenwissen in Form von Scores und Aufgabenanalysen verknüpft. Zusammenfassend werden die folgenden zwei explorativen Forschungsfragen formuliert:

FF1: Welche Fähigkeitsprofile des FDW lassen sich in den Score-Daten aus ProfiLe-P+ mithilfe von Clusteranalysen finden?

FF2: Durch welchen Sprachgebrauch im Testinstrument zeichnen sich die Personengruppen zu den Fähigkeitsprofilen aus?

Stichprobe und Design

Der vorliegende Datensatz umfasst 846 Bearbeitungen - primär von Physik-Lehramtsstudierenden - des Testinstruments zur Erfassung des FDW in den vier Facetten *Schülervorstellungen, Instruktionsstrategien, Experimente* und *Fachdidaktische Konzepte* nach Gramzow (2015). Nach Ausschluss unvollständiger Bearbeitungen blieben $N = 779$ Bearbeitungen für die Analyse.

Inspiriert von den datengetriebenen Ergebnissen der oben genannten Scale-Anchoring-Analysen des FDW (Schiering et al., 2023; Zeller et al., 2022) wurden die Aufgaben des Testinstruments einer Anforderungsanalyse, angelehnt an die Anforderungsdimensionen nach Anderson und Krathwohl (2001), unterzogen. Dabei wurden aus der theoretischen Modellierung die Dimensionen *Erinnern / Verstehen, Anwenden, Analysieren, Evaluieren* und *Kreieren* abgeleitet. Eine weitere Dimension *Unterrichtssituation* wurde induktiv ergänzt. Eine Aufgabe kann mehreren Anforderungsdimensionen zugeordnet werden. Das Ziel dieser Analyse war es, die Dimensionalität des Datensatzes derart zu reduzieren und somit inhaltlich zu verdichten, dass eine Interpretierbarkeit (Nelson, 2020) der explorativen Analyse ermöglicht wird.

Für die Bearbeitung der zweiten Forschungsfrage wurden die Antworten der Proband:innen zu den einzelnen Aufgaben digitalisiert und personenweise zusammengefasst. Darüber hinaus wurden einige standard NLP-preprocessing Schritte wie *lowercasing, stemming* und *stopword removal* angewendet.

Methodik

Zur Ermittlung des prototypischen Antwortverhaltens von Personenclustern wurde das k-Means (MacQueen, 1967) Verfahren gewählt, da es ermöglicht, Zentrumsvektoren und somit typische Ausprägungen der Scores in den Anforderungsdimensionen für die Cluster zu extrahieren. Um eine gute Abdeckung der Varianz bei gleichzeitig angemessener inhaltlicher Verdichtung zu erreichen, wurde ein Modell mit $k = 4$ Clustern gewählt und an die Daten angepasst. Andere Methoden (wie Latente Klassenanalysen) wurden ebenfalls getestet, ließen sich aber auf den vorhandenen Datensatz nicht anwenden.

Zur Untersuchung der zweiten Forschungsfrage wurde ein Structural Topic Model (Roberts et al., 2019) an die von den Probanden gegebenen Antworten auf die offenen Fragen des Testinstruments erstellt. Ein Topic Model modelliert mathematisch Text als Mix aus Themen und Themen als Mix aus Worten. Ein Structural Topic Model ermöglicht darüber hinaus, Kovariaten (hier die Zuordnung zu den K-Means Score-Clustern) in die Modellierung einfließen zu lassen und darüber hinaus den Zusammenhang zwischen der Fokussierung auf ein Thema mit den Kovariaten in Bezug zu setzen. Um eine Balance zwischen Interpretierbarkeit und Spezifität der Themen zu finden wurden unterschiedliche Modelle erstellt und anschließend ein Modell mit 6 Themen gewählt. Die Themen wurden anschließend auf Basis ihrer charakteristischen Terminologie (unterschiedliche Metriken, siehe Roberts et al., 2019) gelabelt.

Ergebnisse und Ausblick

Das zentrale Ergebnis der Cluster-Analyse (FF1) sind die Ausprägungen der Dimensionen der Clusterzentren (Darstellung in einem Netzdiagramm in Abb. 1). Das grüne und das blaue Cluster weisen hierbei keine signifikanten Unterschiede im Gesamtscore oder

Studienfortschritt auf, diese Unterscheidung stellt also eine nicht-hierarchische Beobachtung dar. Es zeigt sich, dass das das grüne Cluster offenbar deutliche Stärken im Anwenden von FDW und der Kreation von Unterrichtselementen (z. B. Experimente, Material, beschriebene Situation) aufweist, währen das blaue Cluster Stärken in der Analyse und Evaluation von Unterrichtselementen zeigt.

In Bezug auf möglicherweise prototypischen Sprachgebrauch der Score-Cluster (FF2) wurden die Effekte, die die Zuordnung zu einem Score-Cluster auf die Anteile eines Topic-Model Themas in den Testantworten hat, bestimmt. Dabei zeigten sich in der Tendenz Unterschiede zwischen dem evaluierenden und kreativen Cluster: Währen sich das kreative Cluster eher auf die Nutzung und Begründung von Beispielen[1] fokussiert, konzentriert sich das evaluierende Cluster eher auf Schülervorstellungen[2]. Dabei kommt im Umkehrschluss die tiefgründigere Auseinandersetzung mit den gewählten Beispielen im kreativen Cluster eher etwas zu kurz, während dem evaluierenden Cluster teilweise die Erzeugung geeigneterer Beispiele schwerer fällt.

Insgesamt lassen sich somit in dieser Analyse auch nicht hierarchische Fähigkeitsprofile (Cluster & prototypischer Sprachgebrauch) identifizieren, auch wenn diese Einordnung eher „Tendenzausprägungen“ als echte „latente Gruppen“ darstellen. Die Ergebnisse können zukünftig für informativeres inhaltliches Feedback für Proband:innen oder für formative Diagnostik zur Gestaltung von Lehrveranstaltungen genutzt werden. Um diese Ziele leichter zugänglich zu machen, soll im nächsten Schritt des Projekts die Auswertung des Testinstruments mithilfe weiterer NLP- und ML-Methoden automatisiert werden.

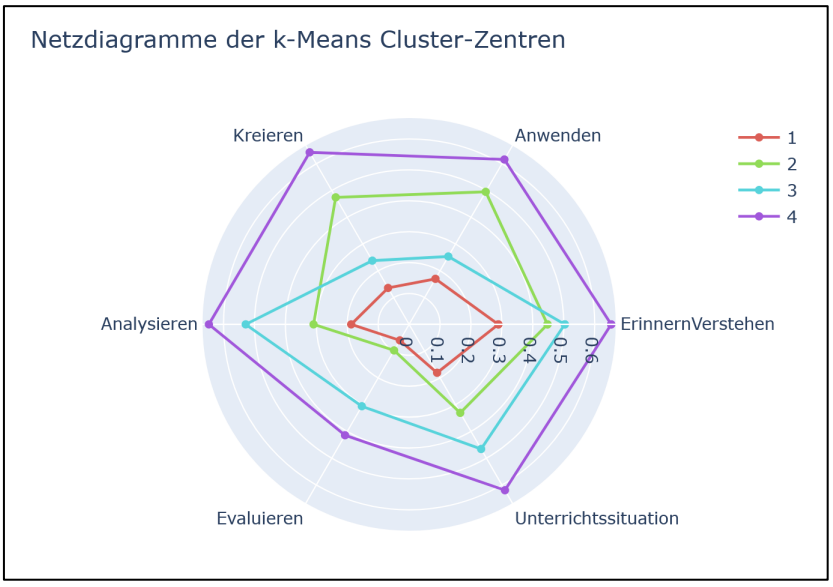

Abb. 1: Netzdiagramme der Clusterzentren.

[1] Begriffe wie „Auto“, „Berg“, „Anschauungsmaterial“, „Veranschaulichung, „Denkanstoß“, „zeigen“
[2] Begriffe wie „Schülervorstellung“, „Alltagserfahrung“, „kognitiv“, „Konflikt“, Begriffswechsel, „umdeuten“

Literatur

Alonzo, A. C., Barendsen, E., Berry, A., Borowski, A., Carpendale, J., Kam Ho Chan, K., Cooper, R., Friedrichsen, P., Gess-Newsome, J., Henze-Rietveld, I., Hume, A., Kirschner, S., Liepertz, S., Loughran, J., Mavhunga, E., Neumann, K., Nilsson, P., Park, S, Rollnick, M . . . Wilson, C. D. (2019). The Refined Consensus Model of Pedagogical Content Knowledge in Science Education. In A. Hume, R. Cooper & A. Borowski (Hrsg.), *Repositioning Pedagogical Content Knowledge in Teachers' Knowledge for Teaching Science* (S. 77–94). Springer Singapore. https://doi.org/10.1007/978-981-13-5898-2_2

Anderson, L. W., & Krathwohl, D. R. (Hrsg.). (2001). *A taxonomy for learning, teaching, and assessing A revision of Bloom's taxonomy of educational objectives* (4. Aufl.). Longman. Carlson, J., Daehler, K. R.,

Gramzow, Y. (2015). Fachdidaktisches Wissen von Lehramtsstudierenden im Fach Physik: Modellierung und Testkonstruktion. In H. Niedderer, H. Fischler & E. Sumfleth (Hrsg.), *Studien zum Physik- und Chemielernen* (Bd. 181). Logos Verlag.

Hattie, J., & Timperley, H. (2007). *The Power of Feedback. Review of Educational Research*, 77(1), 81–112. https://doi.org/10.3102/003465430298487

Jordans, M., Zeller, J., Große-Heilmann, R. I., & Riese, J. (2022). Weiterentwicklung eines physikdidaktischen Tests zum Online-Assessment. In S. Habig (Hrsg.), *Unsicherheit als Element von naturwissenschaftsbezogenen Bildungsprozessen, Tagungsband der GDCP Jahrestagung 2021*. Gesellschaft für Didaktik der Chemie und Physik.

Kaiser, G., Bremerich-Vos, A., & König, J. (2020). Professionswissen. In C. Cramer, J. König, M. Rothland & S. Blömeke (Hrsg.), *Handbuch Lehrerinnen- und Lehrerbildung* (S. 811–818). Klinkhardt. https://doi.org/10.35468/hblb2020-100

Kröger, J. (2019). *Struktur und Entwicklung des Professionswissens angehender Physiklehrkräfte* [Diss., Christian-Albrechts-Universität Kiel].

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam & J. Neyman (Hrsg.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (S. 281–297, Bd. 1). University of California Press.

Nelson, L. K. (2020). Computational grounded theory: A methodological framework. *Sociological Methods & Research*, 49(1), 3–42. https://doi.org/10.1177/0049124117729703

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). stm: An R Package for Structural Topic Models. *Journal of Statistical Software*, 91(2), 1–40. https://doi.org/10.18637/jss.v091.i02

Schiering, D., Sorge, S., Keller, M. M., & Neumann, K. (2023). A proficiency model for pre-service physics teachers' pedagogical content knowledge (PCK)—What constitutes high-level PCK? *Journal of Research in Science Teaching*, 60(1), 136–163. https://doi.org/10.1002/tea.21793

Shulman, L. S. (1986). Those Who Understand: Knowledge Growth in Teaching. *Educational Researcher*, 15(2), 4–14. https://doi.org/10.3102/0013189X015002004

Shute, V. J. (2008). Focus on Formative Feedback. Review of Educational Research, 78(1), 153–189. https://doi.org/10.3102/0034654307313795

Vogelsang, C., Borowski, A., Buschhüter, D., Enkrott, P., Kempin, M., Kulgemeyer, C., Reinhold, P., Riese, J., Schecker, H., & Schröder, J. (2019). Entwicklung von Professionswissen und Unterrichtsperformanz im Lehramtsstudium Physik–Analysen zu valider Testwertinterpretation. *Zeitschrift für Pädagogik*, 65(4), 473–491. https://doi.org/10.25656/01:23990

Zeller, J., Jordans, M., & Riese, J. (2022). Ansätze zur Ermittlung von Kompetenzniveaus im Fachdidaktischen Wissen. In S. Habig (Hrsg.), *Unsicherheit als Element von naturwissenschaftsbezogenen Bildungsprozessen, Tagungsband der GDCP Jahrestagung 2021*. Gesellschaft für Didaktik der Chemie und Physik.