Das Schülerinventar zur Beurteilung von Schülertexten

Stefan Keller, Jens Möller

Das Forschungsprojekt ASSET

Die Autoren stellen ein Forschungsprojekt vor, in dem die Kognitionen von Lehrkräften untersucht werden, die den komplexen Urteilsprozessen fachlicher Leistungen zu Grunde liegen.

Dr. Jens Möller ist Professor am Institut für Pädagogisch-Psychologische Lehr- und Lernforschung der Christian-Albrechts-Universität zu Kiel. E-Mail: jmoeller@ipl.uni-kiel.de

Pof. Dr. Stefan Keller ist Stellvertretender Direktor am Institut für Bildungswissenschaften (IBW) der Universität Basel. E-Mail: stefan.keller@fhnw.ch

„Interessanterweise urteilen auch die Lehramtsstudierenden noch strenger als der E-Rater©.“

In unserem Forschungsprojekt gehen wir von der Annahme aus, dass wir an den Schulen über fachlich und pädagogisch gut ausgebildete Lehrkräfte verfügen, die nach wie vor den Schlüssel für den Unterrichtserfolg darstellen. Genau an diesem Gedanken setzt unser Projekt an, die Lehrpersonen in einem Kernbereich der pädagogischen Professionalität (Baumert und Kunter, 2006) auszubilden, der sogenannten Diagnostischen Kompetenz (DK). DK kann man verstehen als Fähigkeit, Personen in ihren Leistungen zutreffend zu beurteilen (Schrader, 2013). Das ist wichtig, weil es zum einen darum geht, faire Noten zu vergeben und objektive Urteile zu fällen. Zum anderen ist es wichtig für die Unterstützung des Lernens zu erkennen, wo eine Schülerin oder ein Schüler steht, um den Unterricht adaptiv anpassen zu können (Südkamp, Kaiser und Möller, 2012). Das sind für uns die zentralen Elemente der Diagnosekompetenz. Wissenschaftlich wird DK meist verstanden als Maß der Übereinstimmung zwischen Lehrerurteilen in der Praxis und objektiv gemessenen Schülerleistungen (Schrader, 2013). Und in diesem Zusammenhang spielt die Künstliche Intelligenz eine bedeutende Rolle, weil mit den dort entwickelten Tools ganz neue Möglichkeiten entstehen, Objektivität bei den Urteilen zu komplexen Schülerleistungen zu erzielen, gehärtete Expertenurteile (sogenannte *Benchmarks*) herzustellen, an denen Lehrkräfte ihre diagnostischen Kompetenzen schulen können.

Die folgenden Ausführungen sind für den Sprachunterricht relevant, aber auch für andere Fächer wichtig. In allen Fächern stehen Lehrpersonen vor der Aufgabe, anhand authentischer Schülerprodukte nach wissenschaftlich geprüften Kriterien objektive Urteile zu fällen. Um dies zu lernen, kann es hilfreich sein, die eigenen Urteile mit den Urteilen von Expertinnen und Experten vergleichen zu können, und zwar solche Urteile, die mit Hilfe künstlicher Intelligenz zusätzlich noch einmal Objektivität erhalten haben. Lehrkräfte erhalten damit die Chance, mögliche Verzerrungseffekte, Urteilstendenzen zu erkennen und sich davor zu schützen.

Solche Studien sind natürlich nicht neu, es gab sie bereits in den 1960er Jahren. Die Forschung zeigt, dass die Urteile der Lehrkräfte von leistungsirrelevanten Informationen beeinträchtigt werden (Birkel & Birkel, 2002; Kaiser et al., 2016). So zeigte bereits Weiss (1965) in der seinerzeit Aufsehen erregenden Untersuchung zur Aufsatzbeurteilung von Lehrkräften, dass identische Lernertexte mit Noten beurteilt wurden, die fast über die komplette Notenskala streuten. Birkel und Birkel (2002) replizierten diese Ergebnisse experimentell und fanden zusätzlich einen Einfluss der Rechtschreibung auf die Beurteilung von Schülertexten. Dieser Einfluss der Rechtschreibung auf das Gesamturteil eines Textes kann dazu führen, dass auch inhaltliche Aspekte des Lernertextes als schlecht bewertet werden und der Schüler insgesamt als weniger intelligent eingeordnet wird. Dem liegt zugrunde, dass ein Merkmal (die Rechtschreibung) so sehr alle anderen Eigenschaften eines Textes (oder der Person) überstrahlen kann, dass das Merkmal einen viel zu großen Einfluss auf die Beurteilung an-

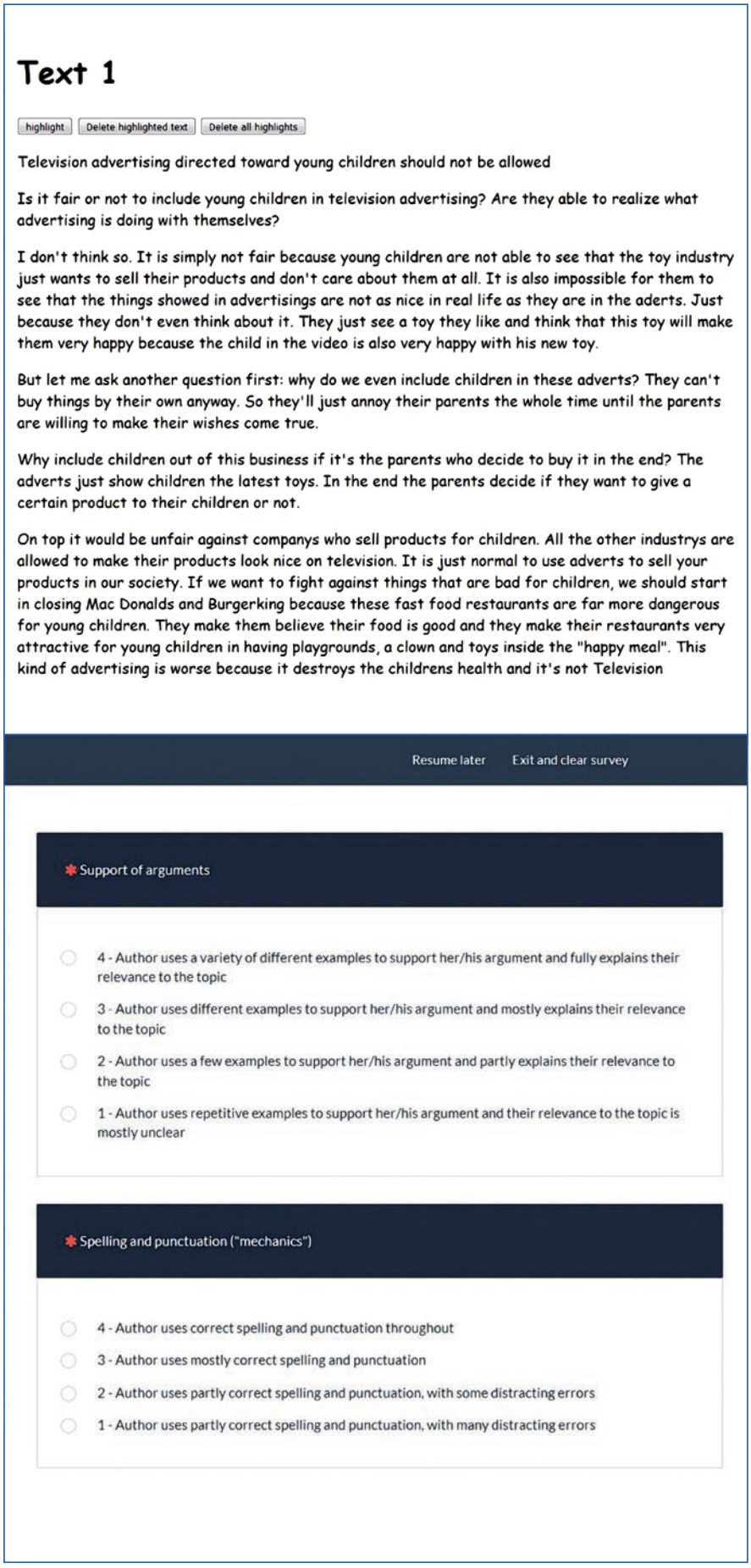

Abbildung 1: Beispieltext und Beurteilungsskala aus dem Projekt MEWS

derer Eigenschaften des Textes (oder der Person) hat. Solche Effekte nennt man Halo-Effekte (Thorndike, 1920). Wie solche Effekte entstehen, das möchten wir mit unserem Projekt ASSET (*Assessing Students' English Texts*) untersuchen.

ASSET funktioniert auf die Weise, dass unsere Probandinnen und Probanden – sowohl erfahrene Lehrkräfte als auch Lehramtsstudierende – in einem Online-Tool die Rolle der Lehrperson übernehmen (s. dazu Jansen, Vögelin, Machts, Keller & Möller, im Druck). „Das Schülerinventar ASSET zur Beurteilung von Schülerarbeiten im Fach Englisch: Drei experimentelle Studien zu Effekten der Textqualität und der Schülernamen. *Psychologie in Erziehung und Unterricht.*" Sie lesen authentische Schülertexte, in diesem Fall aus dem Projekt MEWS (*Measuring English Writing at Secondary Level*), das im Beitrag von Olaf Köller genauer erläutert wird (vgl. dazu auch Keller, 2016). Es handelt sich dabei um argumentative Englischaufsätze aus der 11. Klasse des Gymnasiums. Im Rahmen der MEWS Studie wurden alle diese Texte von zwei speziell geschulten menschlichen Ratern sowie von einer Software zur automatischen Aufsatzbeurteilung (sog. E-Rater©) beurteilt (Ramineni, Trapani, Williamson, Davey und Bridgeman, 2012). E-rater© wurde vom Educational Testing Service (ETS) entwickelt und evaluiert. Wichtige Kategorien von Aufsatzqualität, die E-rater© erfasst, sind Grammatik, Sprachgebrauch, Sprachmechanik, Stil, Organisation und Textaufbau (Ramineni et al., 2012). Es handelt sich also um eine künstliche Intelligenz, welche das menschliche Texturteil nachbildet, dabei aber immun ist gegen typisch menschliche Urteilsfehler wie Müdigkeit, Reihenfolgeeffekte, Tendenzen zu übermäßiger Strenge oder Milde oder Halo-Effekte. Man kann also davon sprechen, dass die Urteile zu den MEWS Texten mit künstlicher Intelligenz objektiviert oder ‚gehärtet' sind.

In ASSET sollen die Teilnehmenden diese Texte beurteilen, einerseits holistisch gemessen im Gesamturteil, aber auch in Bezug auf spezifische Kriterien wie Organisation, Qualität der Argumentation, Sprachmechanik, lexikalische Komplexität, Grammatik, usw. Diese Beurteilung wird nachher mit dem Expertenurteil verglichen und so die di-

agnostische Kompetenz der Teilnehmenden gemessen. Ein Beispiel ist in der Abbildung 1 zu sehen. Auf der linken Seite ist ein Schülertext abgebildet und auf der rechten Seite finden sich Kriterien, nach denen die Probanden den Text bewerten sollen. Gemessen wird, wie nahe sie am Expertenurteil liegen und ihnen wird im Anschluss daran eine Rückmeldung gegeben, inwieweit ihnen das gelungen ist.

Unsere Erfahrung ist, dass diese Rückmeldung von den Probanden sehr geschätzt wird und in der Ausbildung zu sehr fruchtbaren Diskussionen führt. Die Diskussionen gehen dann um Fragen wie „Wann ist ein Essay gut?", „Wie kommt der Experte zu seinem Urteil?", „Wie entsteht der Unterschied zwischen mir und dem Expertenurteil?", „Welche Kriterien sind für die Beurteilung eines solchen Textes relevant (und welche nicht)?" oder „Wie sollen verschiedene Kriterien gewichtet werden?". In solchen Diskussionen lernt man, sein eigenes Urteil an objektiv ‚gehärteten' Expertenurteilen zu eichen, sodass man dann im Sinne der Bildungsgerechtigkeit im Unterricht fachlich eine gute Arbeit leisten kann.

Wie bewerten Lehrkräfte und Studierende im Vergleich zum e-rater©?

In der folgenden Studie wurde gemessen, wie gut Lehramtsstudierende und erfahrene Lehrkräfte Schülertexte im Vergleicht mit e-Rater© beurteilen können. Das System funktioniert in der Weise, dass wir vier Texte aus MEWS ausgewählt haben, welche alle die gleiche Länge haben. Alle waren zuvor mit Hilfe von e-Rater© und menschlichen Ratern übereinstimmend bewertet worden, so dass ein objektiviertes Urteil aus einer Kombination von menschlicher und künstlicher Beurteilung zu jedem Text vorliegt.

Wir (Jansen, Vögelin, Machts, Keller & Möller, i. V.) haben diese bereits bewerteten Texte 47 Lehramtsanwärterinnen und Lehramtsanwärtern vorgelegt sowie auch 36 erfahrenen Englischlehrkräften, die im Durchschnitt 16 Jahre Berufserfahrung hatten. Ein Großteil der Befragten war weiblichen Geschlechts.

Die vier vorgelegten Texte unterscheiden sich in Bezug auf ihre Gesamtqualität. Es wurden jeweils zwei qualitativ hochwertige Texte und zwei qualitativ weniger gute Schülertexte ausgewählt. Jeder Proband hatte die Aufgabe, alle vier Texte zuerst zu lesen, auf diese Weise wurde ein gemeinsamer Bezugsrahmen hergestellt. Würde man Lehrkräften lediglich einen Text zu lesen geben, wäre das aus dem Grunde unfair, weil sie keinen Bezugsrahmen hätten und nicht wüssten, wie sie ihn verankern sollten. Danach konnten die Texte in beliebiger Reihenfolge bezüglich der Gesamtbewertung und der Detailskalen beurteilt werden.

Die Auswertung der Daten der Gesamtbewertung ergab zunächst einen statistisch bedeutsamen Effekt der Textqualität: Insgesamt wurden die schwächeren Texte negativer als die besseren Texte bewertet. Wichtiger für unsere Fragestellung nach der Urteilsqualität von Studierenden und erfahrenen Lehrkräften ist der Befund, dass die erfahrenen Lehrkräfte deutlich strenger bewerteten als die Lehramtsstudierenden, und zwar sowohl in Bezug auf die schwächeren Texte als auch in Bezug auf die stärkeren Texte. In der Konsequenz bedeutet dies, dass die erfahrenen Lehrkräfte weiter von den Expertenurteilen und dem Urteil aus dem E-Rater© urteilen. Werden diese als „wahre" Bewertung bezeichnet, kommt man zu dem Ergebnis, dass erfahrene Lehrkräfte ungenauer urteilen als Lehramtsstudierende. Interessanterweise urteilen auch die Lehramtsstudierenden noch strenger als der E-Rater©.

Für die Detailskalen vergleichen wir nur die Studierenden und die erfahrenen Lehrkräfte, da keine Werte aus dem E-Rater© vorliegen. Auf fast allen Skalen wie beispielsweise Organisation, Qualität der Argumentation, Sprachmechanik und lexikalische Komplexität waren die erfahrenen Lehrkräfte strenger als die Studierenden, nur bei der Beurteilung der Grammatik gab es keine Unterschiede zwischen beiden Gruppen.

Created: September 1, 2012 **Coh-Metrix 3.0** Last updated: Aug. 16, 2017

Enter your input

New technology, such as Smartphones or televisions are ruling our everyday life! Technology is everywhere nowadays. The first thing you do in the morning is get a coffee from your espresso machine, turn on the radio and take a shower. But what if all these things suddenly stop working? I believe that, people should cut down on their use of technology and learn to survive without it! A big plus of having technology doing everything for you, is that you have a lot more time. But how do we use it? Most people go home and watch the news or play video games and they would be bored if we suddenly lost all those things. Even the trip to the grocery store, wich is a privilege to have in the first place, could not be done without cars in most countries. Personaly I could be pretty bored in the evenings, if we did not have all these entertaining technologies or even electricety.

Type text in the image

Save Data

Number	Label	Label V2.x	Text	Full description
Descriptive				
1	DESPC	READNP	7	Paragraph count, number of paragraphs
2	DESSC	READNS	28	Sentence count, number of sentences
3	DESWC	READNW	471	Word count, number of words
4	DESPL	READAPL	4	Paragraph length, number of sentences in a paragraph, mean
5	DESPLd	n/a	2.309	Paragraph length, number of sentences in a pragraph, standard deviation
6	DESSL	READASL	16.821	Sentence length, number of words, mean
7	DESSLd	n/a	9.970	Sentence length, number of words, standard deviation
8	DESWLsy	READASW	1.590	Word length, number of syllables, mean
9	DESWLsyd	n/a	0.932	Word length, number of syllables, standard deviation
10	DESWLlt	n/a	4.907	Word length, number of letters, mean
11	DESWLltd	n/a	2.515	Word length, number of letters, standard deviation
Text Easability Principle Component Scores				
12	PCNARz	n/a	-0.344	Text Easability PC Narrativity, z score
13	PCNARp	n/a	36.690	Text Easability PC Narrativity, percentile
14	PCSYNz	n/a	0.132	Text Easability PC Syntactic simplicity, z score
15	PCSYNp	n/a	55.170	Text Easability PC Syntactic simplicity, percentile
16	PCCNCz	n/a	1.412	Text Easability PC Word concreteness, z score

Abb. 2: Variation von Wortschatz-Qualität

Zusammengefasst lässt sich aus dieser Untersuchung ableiten, dass

- mittels künstlicher Intelligenz „gehärtete" Expertenurteile sich als Vergleichsmaßstab nutzen lassen,
- alle Teilnehmenden zuverlässig zwischen besseren und schwächeren Texten unterscheiden konnten, und zwar sowohl bezüglich des Gesamturteils als auch der Detailskalen,
- erfahrene Lehrkräfte strenger als Experten, Studierende und der E-Rater© bewerten.

Wie kann die Strenge der Lehrerurteile erklärt werden? Eine mögliche Erklärung hat mit dem unterschiedlichen Professionswissen der Gruppen zu tun: Wir haben im Zusammenhang mit dieser Untersuchung auch das Fachwissen der Probanden abgefragt und einen kleinen Englischtest machen lassen – die erfahrenen Lehrkräfte wussten deutlich mehr als die Studierenden, die Ergebnisse zur Strenge der Lehrerurteile blieben aber konstant, wenn das unterschiedliche Wissen der beiden Gruppen statistisch kontrolliert wurde. Die Haupterklärung, die sich für uns aus diesen Untersuchungen ergeben hat, besteht in einem blinden Fleck der Lehrkräfte, einem sogenannten „expert blind spot". Wenn eine langjährige Berufserfahrung und ein profundes Wissen vorliegen, kann die Diskrepanz zwischen den Erwartungen an die Leistungsfähigkeit der Schülerinnen und Schüler und deren tatsächliche Leistungen auseinanderdriften, so dass eine Enttäuschung der Erwartungen zu strengeren Urteilen führt.

Manipulation von Textmerkmalen mit Künstlicher Intelligenz

In der zweiten Studie, die wir hier vorstellen, haben wir die Schülertexte bezüglich einer spezifischen Qualität variiert, nämlich der Qualität des Wortschatzes. Das Ziel dabei ist, den Einfluss des Merkmals „Wortschatz" auf die Beurteilung anderer Merkmale zu untersuchen, also einen Halo-Effekt des Wortschatzes auf Merkmale wie Grammatik, Argumentationsqualität oder Organisation, deren Bewertung eigentlich nicht durch den Wortschatz beeinflusst sein sollte. Die Anlage der Studie ist praktisch identisch mit der ersten Studie. Wiederum haben wir aus einem Korpus von authentischen argumentativen Englischaufsätzen (Gymnasium, 11. Klasse) vier Texte ausgewählt, jeweils zwei Texte hoher Qualität und zwei Texte niedrigerer Gesamtqualität. Zusätzlich haben wir bei jedem dieser vier Texte die Qualität des Wortschatzes variiert, und dabei wiederum auf die Hilfe von künstlicher Intelligenz zurückgegriffen. Den Probanden wurden immer vier Texte gezeigt, die alle Kombinationen realisierten (hohe Qualität/starker Wortschatz; hohe Qualität/schwacher Wortschatz; niedrige Qualität/starker Wortschatz/niedrige Qualität/schwacher Wortschatz).

In der Abbildung 2 ist zu sehen, wie wir die Wortschatzqualität variiert haben. Links ist einer der Schülertexte zu sehen, die die Probanden beurteilen sollen. Um die Wortschatzqualität zu variieren, veränderten wir einzelne Begriffe, tauschten beispielsweise einfache (in Schülertexten vielfach genutzte Wörter) durch solche aus, die seltener in englischen Schülertexten vorkommen und für hochklassige Texte dieses Genres besonders maßgeblich sind. Anschließend wurde der modifizierte mit dem ursprünglichen Text verglichen, indem die Texte verschiedene, auf künstlicher Intelligenz basierende Programme durchliefen. Dies sind Coh-Metrix 3.0 sowie das *Tool for the Automatic Analysis of Lexical* Sophistication (TAALES). Diese Programme erlaubten es uns, empirisch gehärtete Werte für die lexikalische Komplexität sowie die Vielfalt des Wortschatzes in einem Text zu erhalten (zu Details siehe Vögelin, Jansen, Keller, Machts & Möller, 2018). Durch die Verwendung dieser Programme erhalten wir, wie in Abbildung 2 unten zu sehen, einen Hinweis auf die Wortschatzqualität bzw. auf die Komplexität des verwendeten Wortschatzes. Das heißt, wir können überprüfen, ob unsere experimentelle Variation gelungen ist, was in unserem Arrangement Folgendes bedeutet: Wir haben von jedem Text eine Variante mit gutem und eine Variante mit weniger gutem Wortschatz. Ohne die Hilfe künstlicher Intelligenz wären solche Manipulationen ein Stochern im Dunkeln.

Dann ergibt sich dasselbe Vorgehen wie bei der oben dargestellten Untersuchung. Wir haben diese vier Texte unseren Probanden zum Lesen und Beurteilen vorgelegt. Die Forschungsfrage lautete, wie sich die Qualität des Wortschatzes auf die holistische Bewertung auswirkt – was sie ja auch soll, denn schwächerer Wortschatz sollte auch zu einer schlechteren Note führen. Genauso sollte sich die Gesamtqualität der Texte auf die Bewertung auswirken. Dann sollte sich die Qualität des Wortschatzes natürlich auch auf die Bewertung des Wortschatzes auswirken. Allerdings wäre es als Halo-Fehler zu bezeichnen, wenn sich die Qualität des Wortschatzes auf die Beurteilung weiterer Merkmale auswirken würde, die davon nicht betroffen sind, etwa die Qualität der Argumentation oder die Gesamtstruktur des Aufsatzes. Die Ergebnisse zur holistischen Bewertung haben gezeigt, dass die Texte mit dem qualitativ höherwertigen Wortschatz auch zu einer besseren Bewertung geführt haben. Wenn wir den Wortschatz manipuliert haben, konnten wir feststellen, dass die Studierenden das erkannten und die Texte entsprechend positiver bewerteten, wenn der Wortschatz besser war. Entscheidend ist aber, dass wir Halo-Effekte des Wortschatzes auf die Bewertung der Grammatik und der Struktur des Textes fanden. Beide wurden dann negativer bewertet, wenn der Wortschatz weniger komplex war, auch wenn sie eigentlich unverändert waren.

Zusammengefasst:

- Texte mit niedriger Qualität werden auf allen Bewertungsskalen negativer bewertet als Texte mit hoher Qualität.
- Texte mit schwachem Wortschatz werden bezüglich des Wortschatzes negativer bewertet als Texte mit gutem Wortschatz.
- Texte mit schwachem Wortschatz werden auch auf Skalen negativer bewertet, für die der Wortschatz irrelevant sein sollte, z. B. die Gesamtstruktur eines Aufsatzes (Halo-Effekt). Das bedeutet, dass sich einzelne Textmerkmale auch auf Bewertungsskalen auswirkten, die gänzlich unabhängig zu bewerten wären.
- Halo-Effekte ergaben sich auch bei qualitativer Kommentierung durch Lehrkräfte (separate Studie, vgl. Vögelin, Jansen, Keller & Möller, 2018). Halo-Effekte sind also keine Artefakte der verwendeten Skalen, sondern entspringen der Kognition bzw. Wahrnehmung der Texte an sich. Das bedeutet: Auch wenn wir die Probanden haben frei kommentieren

lassen und sie nicht nur per Ankreuzen bewerten mussten, haben wir an den Kommentaren erkennen können, dass unsere Manipulation in einem Merkmal Auswirkungen auf die Bewertung von anderen unabhängigen Merkmalen hatte.

Unser Ziel ist es, diesen Beurteilungstendenzen der Studierenden entgegenzuwirken, und wir entwickeln dazu systematische Maßnahmen. Beispielsweise erproben wir, ob bestimmte Hinweise helfen, diese Fehler reduzieren zu können. Dies kann in der Form eines sog. „Prompts" erfolgen, z. B. „Achten Sie darauf, dass sich die Beurteilung des Wortschatzes leicht auf weitere, davon unabhängige Bereiche der Textqualität auswirken kann. Halten Sie diese Aspekte auseinander".

Möglichkeiten des Schülerinventars ASSET

Das Schülerinventar ASSET ermöglicht es, authentische Schülerleistungen in einem realitätsnahen Setting einzusetzen und bewerten zu lassen. Zugleich ist es in seiner Komplexität eine reduzierte Beurteilungssituation für die Studierenden, weil es sich noch nicht um einen ganzen Klassensatz von Schülerleistungen handelt, es müssen also nicht 25 Schülertexte, sondern nur vier bewertet werden.

Wir variieren relevante Einflussfaktoren auf die Lehrerurteile und können dieses Setting in sehr vielen Fächern einsetzen. Und wenn, wie im Fallbeispiel Englisch, Texte und Instrumentarien zur Verfügung stehen, die sogar noch durch künstliche Intelligenz und maschinelles Rating validiert sind, haben wir sehr gute Möglichkeiten, so etwas wie objektive Urteile als Standard zu verwenden und unseren Probanden Rückmeldungen zu geben, an welchen Stellen sie bereits sehr gut beurteilen und an welchen Stellen sie welche Fehler machen.

Das heißt, der praktische Hauptzweck unserer Forschung besteht darin, die diagnostische Kompetenz der Studierenden zu trainieren und dabei die Vorteile intelligenter Systeme zu nutzen.

Literatur

- Baumert, J.; Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. *Zeitschrift für Erziehungswissenschaften*, 4, 469–520.
- Birkel, P. & Birkel, C. (2002). Wie einig sind sich Lehrer bei der Aufsatzbeurteilung? Eine Replikationsstudie zur Untersuchung von Rudolf Weiss. *Psychologie in Erziehung und Unterricht*, 49 (3), 219-224.
- Jansen, T., Vögelin, C., Machts, N., Keller, S. D. & Möller, J. (im Druck). Das Schülerinventar ASSET zur Beurteilung von Schülerarbeiten im Fach Englisch: Drei experimentelle Studien zu Effekten der Textqualität und der Schülernamen. *Psychologie in Erziehung und Unterricht.*
- Jansen, T., Vögelin, C., Machts, N., Keller, S. D., Köller, O. & Möller, J. (2018). *Who's next to the machine? Comparing teachers' and student teachers' judgments on student essays.* IPL: Kiel University.
- Keller, S. (2016). Measuring Writing at Secondary Level (MEWS). Eine binationale Studie. *Babylonia 3 / 2016*, 46-48.
- Ramineni, C., Trapani, C.S., Williamson, D.M., Davey, T., & Bridgeman, B. (2012). Evaluation of the e-rater® scoring engine for the TOEFL independent and integrated prompts. *ETS Research Report Series, RR-12-06*, i-51.
- Schrader, Friedrich-Wilhelm (2013). Teacher Diagnosis and Diagnostic Competence. *Beiträge zur Lehrerbildung 31* (2):154-65.
- Südkamp, A.; Kaiser, J.; Möller, J. (2012). Accuracy of Teachers' Judgments of Students Academic Achievement: A Meta-Analysis. *Journal of Educational Psychology*, 104(3), 743–762.
- Thorndike, E. L. (1920). A constant error in psychological rating. *Journal of Applied Psychology, 4,* 25–29.
- Vögelin, C., Jansen, T., Keller, S., Machts, N., & Möller, J. (2018). The influence of lexical features on teacher judgements of ESL argumentative essays. *Assessing Writing* 39 (2019), 50-63.
- Vögelin, C.; Jansen, T.; Keller, S. & Möller, J. (2018): The impact of vocabulary and spelling on judgments of ESL essays: An analysis of teacher comments. *The Language Learning Journal.*
- Weiss, R. (1965). *Zensur und Zeugnis. Beiträge zu einer Kritik der Zuverlässigkeit und Zweckmäßigkeit der Ziffernbenotung.* Haslinger.