Zehner, Fabian; Hahnel, Carolin

Artificial intelligence on the advance to enhance educational assessment: Scientific clickbait or genuine gamechanger?

Journal of computer assisted learning 39 (2023) 3, S. 695-702

Quellenangabe/ Reference:
Zehner, Fabian; Hahnel, Carolin: Artificial intelligence on the advance to enhance educational assessment: Scientific clickbait or genuine gamechanger? - In: Journal of computer assisted learning 39 (2023) 3, S. 695-702 - URN: urn:nbn:de:0111-pedocs-318216 - DOI: 10.25656/01:31821; 10.1111/jcal.12810

https://nbn-resolving.org/urn:nbn:de:0111-pedocs-318216
https://doi.org/10.25656/01:31821

Received: 20 February 2023 | Accepted: 30 March 2023
DOI: 10.1111/jcal.12810

EDITORIAL

Journal of Computer Assisted Learning WILEY

Artificial intelligence on the advance to enhance educational assessment: Scientific clickbait or genuine gamechanger?

Fabian Zehner[1,2] | Carolin Hahnel[1,2]

[1]Centre for Technology-Based Assessment (TBA), DIPF | Leibniz Institute for Research and Information in Education, Frankfurt am Main, Germany

[2]Centre for International Student Assessment (ZIB) e.V., Frankfurt am Main, Germany

Correspondence
Fabian Zehner, Centre for Technology-Based Assessment (TBA), DIPF | Leibniz Institute for Research and Information in Education, Frankfurt am Main, Germany.
Email: f.zehner@dipf.de

Abstract

Contributions in the Special Issue: The special issue assembles papers centring around log data analysis, natural language processing, and machine learning used to advance educational assessment. They demonstrate how semi- and unstructured data such as log and text data can, despite their challenging nature, be handled appropriately to benefit educational assessment. In this editorial, we contextualize the special issue's contributions within the diverse field of modern technology-based assessments.

Reflection on Terminology: Moreover, we raise concerns about nowadays' use of the term *artificial intelligence* (AI) in scientific communication. While the contribution of AI to scientific progress is indisputable, the mere use of methods that have evolved within AI research does not necessarily render tools or studies AI-related. We argue that academics have the social responsibility to adopt accurate terminology, given it is integral to scientific rigour and proper scientific communication.

Implications: In view of the inflationary use of the term AI in science, we propose a scheme to locate one's research in the field by focusing on (1) the type of data, (2) the processing involved, and (3) the output of a study and the actions derived from it, which are situated within the (4) scope of a study.

KEYWORDS
artificial intelligence, educational assessment, log data analysis, machine learning, natural language processing

1 | MORE COMPLEXITY FOR EDUCATIONAL ASSESSMENT

As students in schools and universities, or more generally as test takers, we have traditionally been exposed to somewhat dull test formats: bubble sheets, single and multiple-choice, fill-in-the-blanks, or similar formats that mimic the real-world interactions between teachers and their learners rather abstractly. The designed assessment setting is usually rather inauthentic. For us as test designers, such response formats are convenient, mainly for logistical and practical reasons. We rarely consider them explicitly to elicit particular cognitions or behaviour when designing learning and assessment environments. Contemporary assessments, however, put us in a different position. By enabling us to create large or even open spaces for test takers to navigate through and interact with stimulus and learning materials, they allow for more immersive presentation modes and authentic settings; for example, with conversation-based assessment (Yildirim Erbasli, 2022). Moreover, they can incorporate technologies that support the automatic processing of constructed responses (Smith et al., 2021) and enable capturing of log data on test takers' interactions (Goldhammer et al., 2021). While such developments provide more diverse information about test takers and bring us closer to the real-world interactions we seek to study, they also add new complexity to the observations of response behaviour, that is, the resulting empirical data. One way to characterize those more complex types of data is the distinction between semi-structured and unstructured information.

Information is considered structured when the data's 'intended meaning is unambiguous and explicitly represented in [its] structure' (Ferrucci & Lally, 2004, p. 327); or put differently, when it follows a predefined format comprising a narrow range of possible values. That is the case for traditional data consisting of 0s and 1s that represent correct and incorrect responses. Now, think of constructed written or spoken responses as examples of the other pole of this continuum: unstructured information that can take any value—or at least such a high number of values that they are not processable through traditional means of ex-ante specifications of values and their representation. Unstructured information does not convey its meaning through its structure (Ferrucci & Lally, 2004). Unsurprisingly, semi-structured information combines characteristics of both structured and unstructured information with some predefined format that increases in complexity through a high number of possible combinations of structured and unstructured parts. Log data from assessment environments are an example of semi-structured data, as they contain a multitude of information about test takers' interactions with the computer-based environment, including types of events, event descriptors, and their time of occurrence.

The studies of the special issue *Advancing Educational Assessment Through Log Data Analysis, Natural Language Processing, and Machine Learning* centre around data that contain such unstructured or semi-structured information. They use text data, log data, or both in the realm of educational assessment and provide a bouquet of examples of how to tackle the complexity of semi- or unstructured information. Their methods include machine learning, feature engineering, natural language processing and generation, information alignment (e.g. synchronization of different data streams), and other associated methods for mining educational data. *But wait!*, you might think, *this is a list of nowadays common buzzwords—aren't we calling it AI?* And indeed, you are right; they are all commonly referred to as artificial intelligence (AI). So, why are we not calling it by its name? Well, we think that academics have the obligation to critically reflect on their use of terminology, and the concurrent use of the term AI might be troublesome in this respect. In the following sections, we elaborate on why we aim to raise awareness for scientific terminology used in this particular case and on how the special issue's contributions might be framed otherwise in this diverse field.

2 | ARTIFICIAL INTELLIGENCE: MEANINGFUL INNOVATION OR HYPE?

First, let us get the elephant out of the room: AI is not just hype, and it will leave its mark on our society and many research fields. We do not argue with its value. Quite the opposite, the papers of this special issue embrace the methodological approaches that often fall under the umbrella term AI. As a sample of publications on semi- and unstructured data, they demonstrate evidence that many methodological innovations and substantive findings of the field constitute valuable engineering developments, methodological advancements or even breakthroughs, and sometimes meaningful epistemological progress. At their heart lies the challenge to tackle the complexity of representing semi- and unstructured information, creating and processing suitable representations, and aligning (e.g. synchronizing) them in meaningful ways.

Still, although we do not vote against using the term *AI* in general, we intend to raise awareness for the importance of accurate terminology for science communication and the possibility that not every work tagged as AI-related *should* be tagged this way. In other words, while some *methods* and *paradigms* developed in the context of AI research are groundbreaking, *using* them in other contexts does not necessarily render a study or system AI-related. Therefore, we ask: Is AI a genuine game-changer for educational assessment, or is the term used as a clickbait to attract attention?

When the tech industry propagates its products as AI, or even as omnipotent, it obviously follows the goals of its marketing campaigns. In the act of receiving attention at any cost, they exaggerate the power of AI systems and showcase them in celebratory events, such as Google Duplex (Leviathan & Matias, 2018)[1] or Tesla misrepresenting their Autopilot's reliability in an extensively edited video (Jin, 2023). The interests of these companies are commercial in nature. If, in contrast, academic researchers, who serve truth—or at least models that resemble truth—and who are responsible for independently advancing society's knowledge, label their work with the term, they *do* indeed research and develop genuine AI systems and transparently expose limitations, rather than primarily advertising their work. *Or do they?*

Our personal experiences in academia are not always in line with the romanticized view of researchers who serve only to advance scientific knowledge. Let us share an emblematic anecdote. In 2021, the two of us attended an online session at an international educational measurement conference whose title contained the term AI. The chat participation during this session was comparatively lively. So far, our field had primarily distinguished between methods such as machine learning, natural language processing, process mining, neural nets, etc., and had named them accordingly. Thus, it was unexpected that some researchers, on certain occasions, had changed the label of their work, which they have been carrying out for decades, into something vaguer like AI. Seeing this, it seemed to us that *the way* scientific communities like this one have been adopting the term for their agenda lacked some reflection. Therefore, we took the initiative and asked the chat what colleagues thought about attaching a new term to methods that already had their place in the community before. The chat responses were affirmative of our perspective throughout but added one noteworthy comment: *You are right that it's not terribly*

[1]Demo at Google I/O 2018: https://youtu.be/znNe4pMCsD4 [2023-02-14].

accurate. But would the session have been as well-attended as it is now?[2] Baffled by this comment, we could not deny the fact. However, we now have the opportunity to add the questions that logically follow: *Which goal do we prioritize in our scientific work? To attract attention or to be precise by respecting the appropriate use of terminology? How much are we willing to trade one for the other?*

In this editorial, we question the inflationary and often careless use of the term AI and suggest a scheme for how to frame the types of studies or systems presented in this special issue. Before we do so, let us have a look at the term AI next and briefly dive into what it conveyed when it was originally coined at the Dartmouth Summer Research Project, how it developed since, and why we argue that sensible terminology is pivotal for academic researchers.

3 | THE IMPORTANCE OF TERMINOLOGY AND THE AI-TERM

3.1 | Terminology in science

In a well-received paper, Osborne and Patterson (2011, p. 628) stated that 'if a field lacks clarity about the concept that it seeks to explore and promote [...], then it will fail to communicate its meaning and intent to the wider audience'. While the authors overlook the distinction between manifest *terms* symbolizing latent *concepts*, their related *entity in the world* (see Mari, 2015), and (less-specific) *words* (see Teubert, 2004), they make two points that are crucial to our discussion. First, terminology is a matter of science communication. Second, the lack of consensus about a term causes turbulence within and outside the field.

Proper use of terminology is so inherent to science communication that, since the 1930s, there has been a distinct field of research on the analysis and evolution of terms with the ultimate goal of standardization within domains (see Sageder, 2010). For example, the *Joint Committee for Guides in Metrology* publishes the *International Vocabulary of Metrology* for building consensus about the relationships of common terms and their underlying concepts (Mari, 2015). As Schuster (2019, p. 169) states, '[t]erms are a key linguistic tool for the constitution and organization of a scientific domain', thereby distinguishing experts from non-experts through the knowledge and use of such terms. Hence, in this traditional perspective, terms shall be non-arbitrary symbols or linguistic markers that unambiguously communicate knowledge between researchers without a chance of imprecision. Importantly, they also distinguish researchers from people without scientific expertise, who might use the same *word* in a natural way, often with a slightly different meaning.

More recent terminology paradigms emphasize other functions of terms and take a social perspective on terminology. Socio-pragmatic approaches (see Schuster, 2019) stress that authors and speakers equip themselves with terms (Felder & Gardt, 2015), just as humans do with status symbols to signify their membership in a particular group. This social aspect unveils a potential risk of recommending a more careful employment of the AI term. When we suggest to use or avoid certain terms to our academic peers and colleagues—in particular researchers who yet have to find their place in our scientific communities, such as Ph.D. students—this can obviously affect which scientific communities their work is attuned to and might determine which ones they find themselves in at a later stage of their academic career.

[2]Message paraphrased from two subjectively filtered memories.

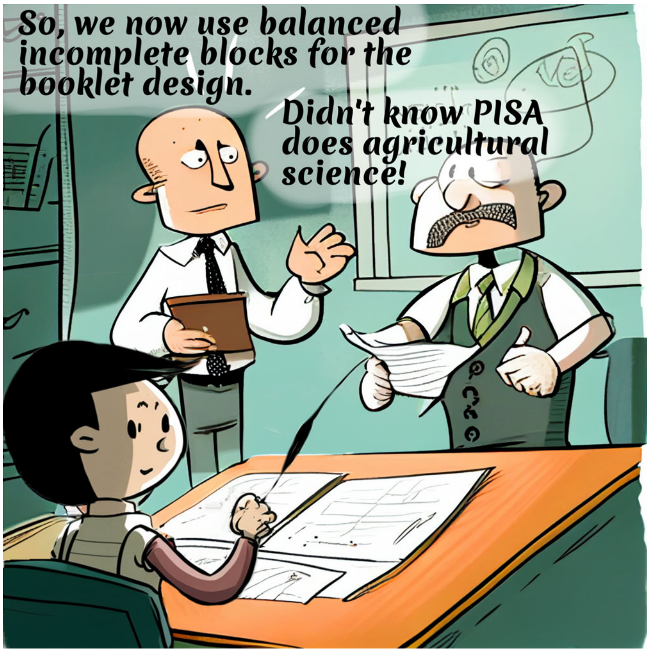

FIGURE 1 On the consequence of using methods outside their original field. Synthetic picture created by Midjourney (v4).

Another important aspect we can draw from the field of terminology is that a term's meanings may naturally change over time (Schuster, 2019). Therefore, it is important to take a step back at this point and briefly recap where the term AI originated, what it referred to, and how its meaning has evolved since.

3.2 | The ever-changing term AI and its usage

The term AI was coined in 1955 by four well-known researchers at the Dartmouth Summer Research Project: John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon (McCarthy et al., 2006). Noteworthy, Minsky had just completed his dissertation on *Neural Nets and the Brain Model Problem* a year prior (Minsky, 1954). In their project proposal, the four researchers formulated the very first purpose of AI research, which should become subject to many changes over the course of the next 70 years: studying intelligent human behaviour by replicating the underlying mechanisms in machines (McCarthy et al., 2006). They selected areas that would later become dedicated fields of research or methods; that is, machines using natural language, forming abstractions, solving problems, and self-improvements (McCarthy et al., 2006). This endeavour very much followed the scheme Richard Feynman would write down

FIGURE 2 Example for (somewhat) hidden inaccuracies in AI output. Synthetic picture created by Midjourney (v4). While the picture is effective at first glance, the inaccurate drawing and placing of hands stands out after closer inspection.

so pointedly to his often-quoted blackboard more than three decades later: 'What I cannot create, I do not understand' (see Way, 2017).

Many different paradigms followed over the decades, driven by sometimes more and often less interest and funding in AI, and even more so with the goal shifting significantly. For example, as opposed to symbolic AI studying the processes *how* humans think, expert knowledge systems introduced the focus on domain knowledge and *what* humans think (Dick, 2019). Fast forward to today, skipping several paradigms and methods, we find ourselves in a new wave of interest from funding organizations, the public, and researchers of almost every discipline. The goal of most research contextualizing itself as AI has utterly left the initial focus on reproducing the mechanisms of human behaviour and shifted now almost exclusively to the reproduction of the behaviour's output (Dick, 2019). That is, technology is now considered artificially intelligent if it reaches the same, or a better, conclusion as humans would do, and not because they work the same way.

This is the critical point of conflict where the boundaries of scientific rigour are regularly pushed. The *way* machines currently come to a conclusion—that is, the methods, such as machine learning or natural language processing—are wholly unbound to the human's mechanisms. As stated above, the dynamic evolution of the term AI is natural and fine. If, however, the methods are used in other contexts, the context itself must relate to the field of AI in order to associate the study with the term. Likewise (see Figure 1), no one would pretend that PISA is agricultural science, even though its booklet design originated from this discipline (Yates, 1936). Back to the case of AI, in other words, the system's or the study's *purpose* must still be in line with the present goal of the term's definition. Probably, most would agree that deep learning applied to weather forecasts is not AI. Then, why would using machine learning (regardless of whether the model is a linear regression, cluster analysis, random forest, or neural network) to build a data-driven model of the relationship between classroom engagement and test performance render the study AI-related? Yet, if a machine learning classifier is designed to replace a task that would usually be carried out by a teacher, then the *purpose* overlaps with today's definition of AI, and thus the use of the term would be more reasonable in this case.

3.3 | Priming the receiver: Terms as linguistic registers

Whether or not to make use of the social function of terms purposefully is a question of the intended effect of the communication. Think about it in the framing of communication between a sender and a receiver. If the sender calls something AI, how does this affect the receiver? As emphasized by socio-pragmatic and socio-cognitive terminology, respectively, the usage of certain terms conveys information about the sender and activates a particular semantic network in the receiver's mind (Schuster, 2019). From a psychological point of view, the sender primes the receiver by naming words that signify the context of the message sent. Similarly, linguistics calls this a register.

The activation of the receiver's pre-knowledge through a linguistic register is helpful but also contains risks. One such risk can be exemplified by the current generation of popular tools based on large language models, such as chatGPT or Midjourney. Their capacity is impressive. So impressive that flaws and inaccuracies can only be detected with some effort, which can result in inaccurate output being used or the disillusionment of users who pay close attention.

Take the Midjourney output in Figure 2 as an example. At first glance, the viewer might admire the expressiveness of the picture and what it conveys on an emotional level. A second look, though, reveals oddities; for example, the improper modelling of hands that might lead to disappointment, disengagement, or even dismissal of the work's quality on the viewer's side, despite its value on other dimensions. Transferred to the use of the AI term, which evokes certain expectations, we as academic researchers and senders may not want to accept these effects on the receivers of our communication: the risk that our research colleagues, funders, or society will disengage from our work because it only borrows from AI-related techniques or even less (e.g. some works relate cluster analysis to AI research, although it did not even originate from this context).

If we accept that some usage of the AI term is problematic and therefore avoid the term in works that do not relate to AI research, how might we replace its function as a register? In the next section, we propose an alternate way for preparing communication receivers to enhance understanding.

4 | TAKING A STEP BACK: A SCHEME FOR FRAMING

Indeed, it has advantages to be able to linguistically use a register to provide the receiver of your communication with context by calling one term. However, especially when a term such as AI may be misleading due to different conceptions of what it encompasses, it is even more purposeful to accomplish the same by simply using more specific terms. That means, we suggest referring to machine learning when this is at the core of a tool or study, to natural language processing or understanding when the main concept is around spoken or written utterances, to data type as being semi- or unstructured, and so on.

In this section, we systematize the components that most studies in the field of the special issue involve, and provide suggestions for alternative term usage and embedding of one's own study. The proposed scheme includes three reciprocal layers that can be regarded as workflow components of a pipeline: (1) raw data, (2) processing and feature extraction, and (3) output and actions. These layers are situated within a (4) study's scope.

(1) **Data** is the foundation of all empirically oriented studies in our field. The data contains raw captured information—in technical terms, at the sensor level—and varies in its degree of technical richness. Traditional educational assessment data consists of raw coded responses, represented by a small range of numbers; in the common extreme case, a dichotomy of correctness. Recent works sometimes contain richer data streams, such as log data from user interactions or natural language from constructed responses. The technically richest data stems from multimodal assessment that can encompass the aforementioned as well as other types of information, such as eye-tracking, functional brain imaging, (neuro-)physiological information, self-reports, contextual information about lighting or temperature, and many more. Figuratively spoken, traditional data is more like the sounds of a digital keyboard, delivering the melody in a rough shape, whereas richer data streams resemble the more resonant tones of a mechanical grand piano, and multimodal streams might correspond to a full-fledged orchestra capable of delivering a highly immersive sensual experience to the audience. Obviously and importantly, the addition of data streams does not automatically result in additional valuable information, just as the addition of random musical instruments does not automatically lead to a more hedonistic experience for an orchestra's audience. Synchronization and weighting of data streams are non-trivial and require a conductor. Moreover, the degree to which the captured information is tied to the actual construct of interest typically decreases from traditional to multimodal data significantly. Information on the assessment process can be worthwhile—for example, in creating evidence on the validity of test score interpretation (Goldhammer et al., 2021)—but do not necessarily need to.

(2) The next layer takes the data as the input for **Processing and Feature Extraction**. Traditional, structured data requires less processing work than semi- and unstructured information. With multimodal streams, synchronization and weighting can be an additional challenge. Representing the information of interest based on the raw data is called feature extraction. This can be done, for example, by means of embeddings if the semantics of natural language data are relevant. The same is true for log events. For example, both natural language and log events can be represented in *n*-grams. This layer can also involve feature engineering to create top-down features, as is often done in the design of process indicators based on log events, feature selection, or both.

(3) The resulting feature values are then often fed into the third layer, **Output and Action**. This generative process can include classification, prediction, decision-making, or other outputs, such as in end-to-end deep learning architectures. While the first layer addresses the data, this third layer is dedicated to the model and its output, with the second intermediate layer specifying the data handling and the operationalization through features. Therefore, this third layer involves model optimization, most often in the form of machine learning algorithms. Based on the output, certain actions can sometimes be derived, such as recommendations to a learner, feedback to a test taker, or suggestions to a teacher.

(4) Finally, and most importantly, all three layers are nested within the centrepiece of a work—its purpose or **Scope**. Does the work address a basic method to be used within a larger environment? Or does it apply a known method to a new context? Is it ready for use, and if so, for whom? The innate degree of automation of a work is also of interest in this context, as some methods and tools are designed for complete automation, while others exclusively assist humans, or even team up human and machine for a hybrid approach.

5 | PAPER CONTRIBUTIONS

We are delighted to present a jam-packed issue with excellent papers from colleagues who demonstrate the relevance and challenges at each level of the proposed scheme. The special issue brings together at least two communities for which the data type is the most distinctive feature. Therefore, we structure the following description of the papers by their primarily used data type. Moreover, we describe the studies' most characteristic features according to the above scheme. Of course, it is up to you, the reader, whether you think it is necessary for any of them to have the AI-badge added on top.

5.1 | Log and text data

There are two contributions in this special issue that link the two worlds of semi-structured and unstructured data (i.e. log and text data) and demonstrate the value of their joint analysis at the output level. Flor and Andrews-Todd (2022) classified actions consisting of log events from learner–environment interactions and learner-generated chat messages during a collaborative problem-solving task. Although they did not directly align the streams of log and chat message data, they used a multi-class ontology with the ultimate goal of replacing human annotation. Their work showed a promising superiority of word embeddings over *n*-grams for the text data at the representational level.

While Flor and Andrews-Todd (2022) built separate classifiers from log and text data, Raković et al. (2023) built one classifier based on log and text data from a writing task based on multiple texts. The classifier is a set of process indicators of self-regulated learning and linguistic product features of the learners' essays. The authors' aim was to predict the quality of learners' writing in this task and to identify the most important predictor variables.

5.2 | Log data

The scope of the papers that focus on the use of log data mostly centres around the additional diagnostic value of specific behavioural patterns. In the case of the paper by Ifenthaler et al. (2023), these behavioural patterns refer to students' engagement with self-assessments that are supposed to help them preparing for their exams over the period of a 12-weeks Bachelor's course. Using k-means clustering and process mining, the authors identify two groups with different navigation behaviour and align this behavioural data with students' self-reported learning strategies.

The two contributions of He et al. (2023) and Hahnel et al. (2023) demonstrate a surprisingly similar approach to identifying behavioural patterns in digital reading. Both applied methods of cluster analysis using a large international student sample and associated the identified cluster with students' reading performance and socio-demographic characteristics. The paper by He et al. (2023) is remarkable in that the authors used the method of dynamic time warping to identify clusters based on navigation and time sequences. The paper of Hahnel et al. (2023), in comparison, used hierarchical clustering based on process indicators from nine digital reading tasks and investigated cluster differences across tasks, languages, and countries.

The paper of Nagy et al. (2023) could be considered the most conventional contribution of the special issue in the field of educational assessment, as it focuses on the comparison and evaluation of different item response models. The authors investigate test disengagement with models that incorporate item response times and contextual information of item presentation (i.e. item position within a test). Their focus was on the output layer, as they tested multiple models with the aim of disentangling different types of disengaged response behaviour that refers to rapid guessing and test-taking persistence.

5.3 | Text data

Somewhat surprisingly, the scope of all papers that addressed text data involves some kind of focus on the human. While Gombert et al. (2023) and Lottridge et al. (2023) aimed at achieving post hoc explainability of classifiers for humans' interpretations, the other three papers directly involved human actors at some point in the envisaged process. Let us start with the explainability-focused papers.

To automatically code concepts presented in short text responses to energy-related questions, Gombert et al. (2023) leveraged fine-tuned transformers and compared their classification performance and explainability with feature-based representations. The authors not only demonstrated remarkable classification performance for a multi-class task with seven codes and the superiority of transformers, but also applied a technique to detect important subwords that predominantly determined the codes. Through this glass-boxing, they were able to uncover shortcuts (i.e. heuristics) learned by the transformers, which were largely inconsistent with the words identified by humans as important.

Although in a very different application context, Lottridge et al. (2023) followed a similar purpose and methodological setup, which led to similar results to those of Gombert et al. (2023) regarding the lack of correspondence between what machines and humans identified at the output level as the determining factors in the classification of responses. The purpose of the built classifier is to identify students at risk—due to anxiety, abuse, suicidal tendencies, and the like—who mention their critical situation in their text responses when participating in a large-scale assessment. The ethical scope of an excellent application of state-of-the-art technology coupled with a focus on explainable methodology makes this paper a valuable contribution to the field.

Patel et al. (2023) contributed a paper that was directly aimed at teachers, too. Focusing on the output and action level, they used the generative large language model GPT-3 to reduce the linguistic difficulty of mathematical word problems. While the generated output showed reduced reading difficulty in common metrics, they encountered severe difficulties in GPT-3's capability to control the mathematical operations involved.

Another paper reporting on a tool with natural language processing capabilities for teachers was provided by Botelho et al. (2023). The authors' goal at the output level was to identify recommendable feedback snippets for short text responses. These snippets had been written by the teacher before when providing feedback on similar responses from other students. In three studies, the author developed a metric to indicate similarity between responses and used the bidirectional transformer Sentence-BERT to automatically score the text responses, which showed no systematic explanations for misclassifications.

Similar to the study by Lottridge et al. (2023), the study by Andersen et al. (2023) was situated in large-scale assessments where large amounts of human resources are devoted to manually coding constructed responses. To keep the human in the loop and in control, Andersen et al. (2023) developed a method that dynamically varies the amount of automatically scored short text responses, depending on the performance of the automatic system. The system reduces the coding burden on the manual coder by propagating codes automatically in the background when a certain level of statistical certainty of accurate automatic coding can be assumed. It leaves the coding of the responses only to the human if its representation of the text responses does not satisfy the required automatic coding performance.

6 | SOCIAL RESPONSIBILITY OF ACADEMIC RESEARCHERS: A CONCLUSION

With its density of highly relevant and successful contributions in one place, the special issue demonstrates that the blending of related fields is fruitful and will soon become increasingly important in the context of multimodal assessment. In this editorial, we argued against the inflationary use of the term AI and recommended the use of terms that specify the actual characteristics of studies and tools. First, if the purpose of a study is not related to AI, but it simply uses a method associated with AI, readers and listeners gain more from knowing whether a study analysed semi- or unstructured data, as well as some important characteristics of the processing, representation, and modelling of the data, than from being exposed to a vague buzzword. Second, when academic researchers communicate science in written publications, talks, or transfer formats, they have the social responsibility to prioritize scientific rigour (i.e. appropriate terminology) when promoting their research through a term that serves as a common buzzword and conflicts with the state-of-the-art language in their field. We presented a three-tiered scheme of components typically found in educational assessment studies that might be associated with what is currently called AI. If the purpose of the work is indeed AI, then labelling it as such is straightforward. However, if the focus is on a single element from any of the three layers, or if a statistical analysis is used that only has been previously used in a machine learning context, the authors of the study might want to consider critically reflecting on the purpose of their self-proclaimed association with AI.

> *The difference between the almost right word and the right word is really a large matter—'tis the difference between the lightning-bug and the lightning.*
> *—Mark Twain*

ACKNOWLEDGEMENTS

It's said that writing is the easy part, but we know better. Big thanks to the authors who poured their hearts and souls into these papers, and to the reviewers who provided the feedback that made them shine. Special thanks to Ben Buckton. After so much talk of terminology, his magical eyes helped ensure that we stayed on the right terminological path. Open Access funding enabled and organized by Projekt DEAL.

FUNDING INFORMATION

No third-party funding involved.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/jcal.12810.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analysed in this study.

ORCID

Fabian Zehner https://orcid.org/0000-0003-3512-1403

Carolin Hahnel https://orcid.org/0000-0003-2394-3944

REFERENCES

Andersen, N., Zehner, F., & Goldhammer, F. (2023). Semi-automatic coding of open-ended text responses in large-scale assessments. *Journal of Computer Assisted Learning*, *39*.

Botelho, A., Baral, S., Erickson, J. A., Benachamardi, P., & Heffernan, N. T. (2023). Leveraging natural language processing to support automated assessment and feedback for student open responses in mathematics. *Journal of Computer Assisted Learning*, *39*.

Dick, S. (2019). Artificial intelligence. *Harvard Data Science Review*, *1*. https://hdsr.mitpress.mit.edu/pub/0aytgrau

Felder, E., & Gardt, A. (2015). Sprache–Erkenntnis–Handeln. In E. Felder & A. Gardt (Eds.), *Handbuch Sprache und Wissen* (pp. 3–33). De Gruyter.

Ferrucci, D., & Lally, A. (2004). UIMA: An architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, *10*, 327–348.

Flor, M., & Andrews-Todd, J. (2022). Towards automatic annotation of collaborative problem-solving skills in technology-enhanced environments. *Journal of Computer Assisted Learning*, *38*(5), 1434–1447.

Goldhammer, F., Hahnel, C., Kroehne, U., & Zehner, F. (2021). From byproduct to design factor: On validating the interpretation of process indicators based on log data. *Large-Scale Assessments in Education*, *9*, 20.

Gombert, S., Di Mitri, D., Karademir, O., Kubsch, M., Kolbe, H., Tautz, S., Grimm, A., Bohm, I., Neumann, K., & Drachsler, H. (2023). Coding energy knowledge in constructed responses with explainable nlp models. *Journal of Computer Assisted Learning*, *39*.

Hahnel, C., Ramalingam, D., Kroehne, U., & Goldhammer, F. (2023). Patterns of reading behaviour in digital hypertext environments. *Journal of Computer Assisted Learning*, *39*.

He, Q., Borgonovi, F., & Suárez-Alvarez, J. (2023). Clustering sequential navigation patterns in multiple-source reading tasks with dynamic time warping method. *Journal of Computer Assisted Learning*, *39*.

Ifenthaler, D., Schumacher, C., & Kuzilek, J. (2023). Investigating students' use of self-assessments in higher education using learning analytics. *Journal of Computer Assisted Learning*, *39*, 255–268.

Jin, H. (2023). Tesla video promoting self-driving was staged, engineer testifies. https://www.reuters.com/technology/tesla-video-promoting-self-driving-was-staged-engineer-testifies-2023-01-17/

Leviathan, Y., & Matias, Y. (2018). Google duplex: An AI system for accomplishing real-world tasks over the phone. https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html

Lottridge, S., Woolf, S., Young, M., Jafari, A., & Ormerod, C. (2023). The use of annotations to explain labels: Comparing results from a human-rater approach to a deep learning approach. *Journal of Computer Assisted Learning*, *39*.

Mari, L. (2015). Evolution of 30 years of the international vocabulary of metrology (VIM). *Metrologia*, *52*, R1–R10.

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the Dartmouth summer research project on artificial intelligence: August 31, 1955. *AI Magazine*, *27*, 12.

Minsky, M. L. (1954). *Neural nets and the brain model problem* [Unpublished dissertation]. Princeton University, Princeton, NJ.

Nagy, G., Ulitzsch, E., & Lindner, M. A. (2023). The role of rapid guessing and test-taking persistence in modelling test-taking engagement. *Journal of Computer Assisted Learning*, *39*.

Osborne, J. F., & Patterson, A. (2011). Scientific argument and explanation: A necessary distinction? *Science Education*, *95*, 627–638.

Patel, N., Nagpal, P., Shah, T., Sharma, A., Malvi, S., & Lomas, D. (2023). Improving mathematics assessment readability: Can large language models help? *Journal of Computer Assisted Learning*, *39*.

Raković, M., Iqbal, S., Li, T., Fan, Y., Singh, S., Surendrannair, S., Kilgour, J., van der Graaf, J., Lim, L., Molenaar, I., Bannert, M., Moore, J., & Gasevic, D. (2023). Harnessing the potential of trace data and linguistic analysis to predict learner performance in a multi-text writing task. *Journal of Computer Assisted Learning*, *39*.

Sageder, D. (2010). Terminology today: A science, an art or a practice? Some aspects on terminology and its development. *Brno Studies in English*, *36*(1), 123–134.

Schuster, B.-M. (2019). The contribution of terminology research to the understanding of science communication. In A. Leßmöllmann, M. Dascal, & T. Gloning (Eds.), *Science communication* (pp. 167–186). De Gruyter.

Smith, E. H., Hollander, J., Graesser, A. C., Sabatini, J., & Hu, X. (2021). Integrating SARA assessment with reading comprehension training in AutoTutor. *English Teaching*, *76*, 17–29.

Teubert, W. (2004). Language as an economic factor: The importance of terminology. In G. Barnbrook, P. Danielsson, & M. Mahlberg (Eds.), *Meaningful texts: The extraction of semantic information from monolingual and multilingual corpora* (pp. 96–106). A&C Black.

Way, M. (2017). What I cannot create, I do not understand. *Journal of Cell Science*, *130*, 2941–2942.

Yates, F. (1936). A new method of arranging variety trials involving a large number of varieties. *The Journal of Agricultural Science*, *26*, 424–455.

Yildirim Erbasli, S. N. (2022). *Conversation-based assessments: Measuring student learning with human-like communication* [Doctoral thesis]. University of Alberta.