Preprint, to appear 2020 in For The Learning of Mathematics

Scaling up qualitative mathematics education research through Artificial Intelligence methods

Christian Büscher

Introduction: a question of scale

The technological advances of recent years have greatly expanded the possibilities of qualitative mathematics education research. With lightweight cameras and virtually unlimited disk space, researchers now have the tools that almost trivialize the collection of large amounts of qualitative data. But frustratingly, in-depth qualitative analysis is so complex that these potentials cannot be utilized. For example, in a four-year study, the author collected 1510 minutes of video data, but due to the need for focus was only able to analyze 450 minutes in detail (Büscher, 2018). It seems that the methods of data analysis have not kept pace with the methods of data collection.

This is indicative of a larger divide in mathematics education research. Although most distinctions of research relate to the difference between *quantitative* and *qualitative* methodologies, this can also be seen as a question of *scale*. Quantitative methodologies are used in *large-scale* research, favoring statistical generalizability over in-depth description. Qualitative methodologies are employed in *small-scale* research, favoring the opposite. Mixed-methods approaches combine both methodologies in an attempt to draw from both strengths, but do so through sub-studies that are themselves either quantitative or qualitative.

Qualitative methodologies are so labor-intensive that they only permit small-scale research. Yet with the right tools, it might well be possible that qualitative methodologies can be scaled up. Today, machine learning and artificial intelligence (AI) might provide just such tools for solving this problem of qualitative research. Recently, researchers have begun investigating the potentials of new AI technology for research practice. For example, Gurevych and colleagues (2018) show how the research workflow in social science could be substantially changed and supported by machine learning technology. In mathematics education research, Kersting and colleagues (2014) investigate the potentials of automated scoring of teacher answers. With the field of AI research rapidly expanding, more insights into the possibilities of using AI methods for scaling up qualitative mathematics education research are needed.

This paper examines these possibilities in two parts: (1) a report on a small study using a basic AI method, the simple neural network, for automated analysis of transcript data. The purpose of this part is to provide first impressions on what is possible using AI methods, how such studies need to be constructed, and on challenges and limitations associated with such studies. Afterwards, (2) the results of the study serve as the foundation for the main point of this paper, a reflection on the methodological potential of AI.

Problems of AI and mathematics education research

The term ‘artificial intelligence’ often evokes images of sentient robots and intelligent supercomputers. However, this is not an adequate description of the field of AI. Instead, AI is better understood as a collection of

various methods that allow computers to perform tasks that are traditionally thought to belong to humans: "Artificial Intelligence is the study of how to make computers do things at which, at the moment, people are better" (Rich, Knight, & Nair, 2009, p.3).

There are numerous 'things' which computers are increasingly able to 'do'. Some of the most common are shown in Table 1. These AI tasks are similar to many tasks performed in qualitative mathematics education research. For example, deductive transcript analysis consists of finding instances of categories within text data, resonating with text analysis. Mathematics education research tasks with a higher degree of interpretation, such as the identification of student reasoning behind actual utterances can be considered tasks of sentiment analysis.

Table 1: Tasks of AI an mathematics education research

Task	AI example	ME research example
Text classification	Given a text, is it an example of a news report, a story, or any other form of text?	Given a teacher utterance, is it a question, an explanation, or an instruction?
Text analysis	Given a text, what are the contents spoken about?	Given a conversation between teachers, what are the contents talked about?
Sentiment analysis	Given a text (a tweet, a post, …), is it (implicitly) inciting aggression or calling for hurting people?	Given a student answer, what are the concepts and conceptions implicit in the words?
Sequence-to-sequence mapping	Given a sentence, what is a sentence expressing the same meaning in another language?	Given a student explanation, what are the warrant, backing, and claim of the argument?
Computer vision	Given an image, what are the objects depicted?	Given a classroom video, what are the gestures employed by the teacher?

Thus, many tasks handled manually by researchers in mathematics education could possibly be supported by automated tools using AI methods. Some researchers already point out possible directions automated analysis could take. For example, Kersting and colleagues (2014) use the machine learning model of naïve Bayes classifiers to score teachers' short written answers on assessments. They show how automated scoring can possibly complement human scoring. Yet much of the potential of machine learning still needs further inquiry. AI methods do not necessarily provide *new* ways of doing research. Moreover, the tasks depicted in Table 1 all relate to the special case of *deductive* analysis. However, if it was possible to use AI methods for deductive analysis, this would greatly improve the *scale* in which some qualitative research could be carried out.

Mathematical content in teachers' talk

One problem for AI are the large amount of data needed as soon as the data become complex and ambiguous – which qualitative data notoriously are. This presents a problem for the methods used here if only a small amount of data are available (for example, anything under 10.000 samples). In order to keep the data as clear as possible, while still producing results interesting to the research community, this study aims at automatically identifying *mathematical content* in teachers' talk.

This focus arose in the context of a teacher development project for inclusive mathematics education. In order to foster learning for all students, teachers need to fulfill teaching jobs such as setting suitable learning goals for their students, which requires careful attention to the mathematical content (Büscher, 2019). The quality of initiated learning processes in the classroom depends less on surface features (e.g., pedagogies) and more on in-depth features such as cognitive demands of the content (Hiebert & Grouws 2007). This makes the mathematics employed by the teachers and the way of employment important research priorities (Hill et al., 2008). Thus, qualitative research on teacher professional development needs tools for uncovering the mathematical content in teachers' talk. If it was possible to automate the task of identifying mathematical content, it would become possible to scale analysis up from single professional development courses to larger development programs.

The task of automated identification of mathematical content in teachers' talk can therefore be considered relevant for research for at least two reasons: (1) it provides a rough but quick quantification of mathematical content, allowing a quick comparison between different courses; and (2) it saves time. This second reason is not only obvious, but rather crucial. Identifying mathematical content itself is not a very demanding task, but it does take a substantial amount of time. An automated analysis could allow researchers to quickly filter out the possibly interesting utterances (i.e. those with mathematical content), which could then provide the basis for a more sophisticated qualitative analysis (e.g. how and when is the mathematical content articulated?) Thus, automated analysis could provide researchers with a tool to focus on more complex qualitative research questions. Should the simple automated identification of mathematical content succeed, more sophisticated AI methods could even provide a more in-depth automated analysis.

A short introduction to simple neural networks

Out of the vast field of AI, *machine learning* is one (still vast) subfield (see Russell & Norvig, 2010 for an in-depth treatment of the field). It deals with problems for which the underlying rules and relationships are either unknown, too complex, or for which no algorithms exist. Instead of starting with an analysis of the problem, machine learning begins with data comprising instances of the problem as well as their answers. The field of machine learning covers approaches for *agents* (computers, robots, programs …) to discover the patterns between problems and their answers in an attempt to establish the underlying rules of the problem. These rules can then be used to produce answers to new instances of the problem in the future.

Simple Neural Networks are one model of machine learning. Although their name implies a model of the brain, this has mostly historical reasons. Their basic function can be explained by an example of identifying mathematical content in teachers' talk. A suitable neural network should be able to distinguish talk about *mathematical* functions and *didactical* functions (e.g. of representations). Figure 1(1) shows such a neural network, represented as a *graph* with weighted, directed edges that only flow into one direction.

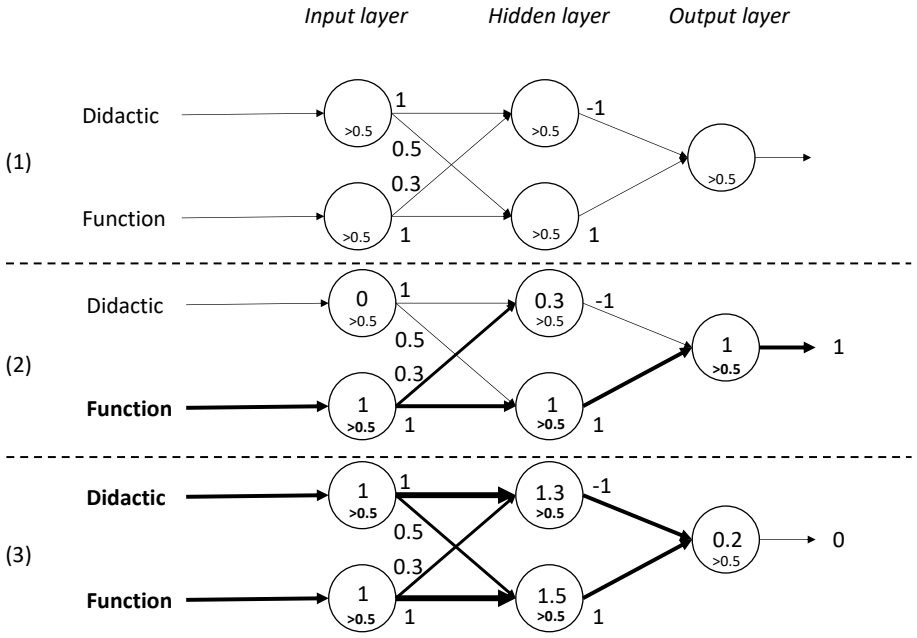

Figure 1: A trained simple neural network identifying mentions of mathematical functions

The nodes are organized in *layers*. Each node of a layer is connected to every node of predecessor and successor layers. Here, the *input* layer consists of only two words 'Didactic' and 'Function'. In practice, input layers are very large, for example consisting of one node for every word in a dictionary. The *output* layer represents the answers computed by the model. In this example, the model should output '1' if a teachers uses the word 'Function', but '0' if it is combined with the word 'Didactic'. Information flows from the input layer over *hidden layers* to the output layer according to the edge weights of the graph: information is passed on through the edges by multiplying it with the corresponding edge weight. Through this procedure, the hidden layers perform the actual calculation of identifying mathematical content. They are 'hidden' not because they cannot be observed, but because in practice, their workings cannot meaningfully be interpreted. Neural network designers generally only consider the number and size of hidden layers. On each hidden layer, the information at each node only gets passed on if its input reaches a certain threshold, according to their *activation function*. This can be interpreted as the model selecting only the most important words for the task at hand and discarding the rest. For example, the activation function in Figure 1 only passes information on if the sum of incoming information is greater than 0.5. If a teacher only utters the word 'function' (2), the output node gets assigned the value '1' and thus produces an output of 1. If a teacher says both words "didactic function", the edge weights cause the last node to receive an input sum of 0.2, which is lower than the activation threshold, and thus produces an output of 0.

In practice, input layers consist of several thousand nodes, and there can be several hidden layers with hundreds of nodes. The task of *training* then consists in finding a weight function that maximizes the likelihood of the desired outcome given the training data. This very rough overview provides a description of only the most basic type of neural network. However, for creating and training such a model, knowledge of high-level concepts suffices. Various ready-made tools and libraries of algorithms can be used that implement the details of machine learning, so that the designer of a neural network can focus mostly on the size and layout of the network. This design work is described below.

The experiment

The data set used for training the simple neural network consists of two hours of transcript data of professional development courses collected across various research projects (e.g. Prediger, Kuhl, Büscher, & Buró, in press). As unit of analysis, single sentences in teachers' utterances were chosen. Thus, the data corpus also includes sentence fragments typical of spoken interactions. The data also required some pre-processing: punctuation was removed and any spoken number or fraction replaced by the special words [NUMBER] and [FRACTION]. The sentences were labeled by the research team by hand with '1' if mathematical content was present, and '0' if not. Mathematical content was identified if a teacher made explicit reference to (a) formal mathematical concepts; (b) informal mathematical concepts (e.g. part, whole); (c) formal mathematical activities (e.g. calculating, multiplying) and (d) informal mathematical activities (portioning, counting). Some interpretation on part of the coders thus was necessary to distinguish between general and mathematical uses, for example between didactic functions and mathematical functions.

During the coding process, the team consensually validated their assignments and created a full coding manual. In total, 1343 sentences were labeled, of which 355 were assigned mathematical content. Out of the 988 sentences without mathematical content, 355 were randomly chosen and used in training to achieve a balance of labeled samples.

The data were then divided into three data sets according to standard machine learning methodology. The *training set* was used to train the simple neural network and consisted of 305 randomly selected samples for each of the 355 labeled sentences, creating a training set of 610 samples in total. Given enough training, however, neural networks tend to *overfit* the training data – their weights get set according to the highly special case of the training data and no longer produce good fit for other data. Because of this, a *validation set* was used to measure the model's performance during training, but did not directly influence the training of the model. The validation set consisted of a 110 randomly selected samples of the training data set. Thus, the actual training was conducted on 500 samples, and the actual design of the network was tuned according to the performance on the validation set. Finally, a *test set* was used as a one-time test to measure the end product's performance.

Since neural networks operate on input vectors of numbers, not words, the labeled sentences had to be *vectorized*. For this, *one-hot encoding* (Russell & Norvig, 2010) was employed: since there were 1511 unique words in the corpus, each sentence was turned into a 1511-dimensional vector with values '0' or '1'. For each vector, a '1' at index *k* indicates that the *k*-th word of the data corpus was present in the sentence. The vectors represent the words of a sentence, but do not encode word order or multiple mentions of words. This method was chosen since approaches that encode word order require more sophisticated AI methods, yet do not necessarily perform better than simple neural networks.

The training of the model was carried out with the *TensorFlow* and *Keras* libraries. After the previous steps of data engineering, the work of the researcher in this step consists of tuning number and size of layers until results are satisfactory. This is an empirical work that comes closer to an exploration than to a rigidly structured procedure. Layers are added, removed, enlarged or narrowed, the model is trained and evaluated, and the resulting fit is noted down. In the end, the best discovered configuration of the model is taken and reported. In this study, the model with the best performance consisted of two hidden layers, each consisting of 16 nodes. The full specification of the model[1] could now be used by other researchers to reproduce the results.

[1] 2 dense hidden layers with 16 units with relu activations, followed by a sigmoid classifier. Optimizer and loss function were the Keras-defaults rmsprop and binary crossentropy. Training was conducted for 3 epochs.

The resulting model was then used to automatically code the test set. Table 2 provides exemplary comparison between output of the model and the original labeling for some samples of the test set. The output of the model is a number between 0 and 1, and can be interpreted as the likelihood that the given sample is labeled with '1', i.e. that it shows mathematical content. Usually, this output is rounded to the nearest integer. In this case, the given samples would all have been coded correctly: The first sentences all refer to mathematical terms (fractions, denominators, parts) or to mathematical activity (decomposing), while the last do not refer to mathematics.

Table 2: Correctly predicted labels by the model

Sample	Translation	Output	Label
Weil der Nenner mir ja immer angibt in wie viele gleich große Teile das unterteilt ist	Because the denominator always gives in how many equal parts it is divided	0.97589725	1
Ich denke so was sagen wie wenn ich das Ganze in mehr Teile zerteile brauche ich mehr kleinere Teile	I think saying how when I decompose the whole into more parts, that I then need more and smaller parts	0.9662741	1
Wie sollen die jetzt aufschreiben was es mit den gleichwertigen Brüchen auf sich hat	How should they write down what equivalent fractions are about	0.9239829	1
Oder irgendwie haben die das ganz toll ausgedrückt	Like somehow they expressed themselves really well	0.28179163	0
Was das bedeutet	What it means	0.26308453	0
Die erste Frage ist ja schon	The first question already is	0.1675178	0

In other cases, the coding failed to predict the actual label, as Table 3 shows.

Table 3: wrongly predicted labels

Sample	Translation	Output	Label
Und ich habe dieses Jahr genau das als Wortspeicher gehabt	And this year I put exactly this into a lexical storage	0.82973784	0
Also diese Gleichwertigkeit ja	Like this equality yes	0.19835638	1

These samples were not coded correctly, but simple neural networks provide almost no possibility for reconstructing the reasons that led to these outputs. Overall, this model performed with 76% accuracy on the test set, meaning that it labeled 76% of the data in the same way as the research team had. With the labelling team it had an interrater reliability on the test set of 0.52 measured by Cohen's kappa, which would commonly be considered a moderate agreement. This is not too bad a result, considering that only a very basic model with a very small data set was used. However, it is also possible that the random selection of training and test data has produced sampling effects that would need to be controlled with more sophisticated methods.

Reflections, hopes, & challenges

The results of this study provide what is called a 'proof of concept' in software development, a working prototype showing that the general idea of using AI methods for qualitative research can be feasible, if only in this case in a very limited domain. This provides an opportunity for further reflections on phenomena encountered in the study.

Learning without insight. Taken alone, the results of the automated coding do not seem that impressive. In the process of labeling the data, the labeling team searched for specific keywords (words denoting mathematical terms and activities), and it is not surprising that a computer program can be used to automatically search for such terms. The problem posed here simply is not such a complex problem that usually motivates machine learning. The important difference is that the team's context knowledge of important mathematical words was not used directly to create a traditional algorithm, but that the model *learned* to focus on certain words all by itself from the labeled data. This means that the general method of training a model from labeled data might also succeed if the labeling criteria are more unclear to the labelers.

***Artificial objectivity*.** The most important use of a trained model is that it can now be used to analyze new data. This seemingly trivial observation has important methodological implications as it relates to questions of reproducible results. One method commonly employed to provide an argument for the quality of a coding instrument is the inter-rater reliability: the degree of agreement between two different raters of the same data. High inter-rater reliability is used to indicate that the results could theoretically be reproduced by others and are therefore somewhat objective. AI-based research could provide a different method, because the trained model can be shared. For example, the model trained for this study, along with the labeled data used for training and the training source code, is publically available at https://ai.cbuescher.eu. Not only is this an argument for the reproducibility of this study's results. It also creates the opportunity for other researchers to use the model for their own research. Trained models can provide a kind of 'artificial' objectivity, in the sense that they provide artifacts that can be used by other researchers to create comparable results.

Context and generalizability. In AI research, a model is ultimately judged by whether it generalizes well. This means that the model should identify the patterns in the training data for accurate predictions, but not in a too narrow way. The patterns used for prediction should be able to be used with other data as well. But the training process still is largely determined by the data used in training; a model can learn to focus on patterns that the research team failed to observe, but it cannot transcend the data. Therefore, the use of trained models in automated analysis needs to be handled carefully. The model trained in this study was trained on data from PD courses focusing on fractions, percentages, language, and inclusive mathematics education. It is bound by this context and is unlikely to perform well for other mathematical subjects or other types of communication. This limitation can likely be overcome by increasing the amount and diversity of training data.

Interpreting black boxes. A challenge that arises from the use of neural networks is their black box character. During the training process, the weights of neural networks are adjusted in a complex way, sometimes even depending on randomization. For simple neural networks, there is no known way to effectively gauge what exactly was learned. It seizes on any apparent pattern found in the data, independent of any pattern intended by the researchers. Thus, automatically analyzed transcripts need to be treated carefully. It might be the case that the model did not actually learn to focus on words denoting mathematical content. It could be that only some few teachers talked about mathematics, and that these teachers all spoke in a particular rhetorical style. If this was the case, the model could simply have learned to identify this style of talking, and it would fail to generalize beyond this data set. However, the fact that the model might have learned to focus on words that the labeling team did not deem important could also be used as a method for generating candidates for important categories: If the model recognized a pattern the team did not, this could be used to uncover theoretical blind spots of the research team.

This problem could potentially be solved in the future. The visualization of what a neural network has learned is a current area of AI research, and some more sophisticated models than simple neural networks do allow better visualization. Yet, the question remains: Can, and should, research be based on black boxes?

Black-box definitions. Again, the black box character could also provide deeper methodological and theoretical implications relating to reproducible results. One way to construct intersubjective theories is to provide closed definitions of central theory elements. Definitions provide standardized ways of interpreting constructs and provide relatable ways of analyzing data, thus securing scientific objectivity. Yet, as the work of Wittgenstein (1953/2003) shows, formal definitions cannot be mapped directly to reality in a clear and non-contradictory way. For even research reports are formulated in natural language, which is 'fuzzy' and ambiguous. AI-based research could drop sophisticated definitions of theoretical constructs in favor of a very hands-on approach: A construct could be defined by what a neural network trained on a (very large) list of examples labeled by experts representing research consensus. Although such a definition would not provide much for theory development, it nevertheless could prove very useful in a practical way for actually finding fitting instances within large bodies of data.

Less test-theoretical tests. The identification of possibly fuzzy constructs in large-scale setting also provides another hope. Large-scale tests such as PISA or TIMSS usually apply narrow definitions of important constructs, such as conceptual knowledge. One possible reason for this are the tools available for evaluating large-scale assessments: Since the method of evaluation is quantitative test theory, large-scale tests are designed for easy coding, which leads to closed-form questions that provide indicators for conceptual knowledge or mathematical literacy. In contrast, many problems in AI concern the processing of texts of natural language. AI-based research could possibly train models that can quickly identify constructs such as conceptual knowledge in written or oral student responses that occur more naturally in daily classroom life. In combination with black-box definitions, this could even lead to easily admissible tests for otherwise hard to quantify mathematical competencies and instruction quality, such as the discursive quality of interactions. Thus, AI-based research could lead to less test-theoretical large scale tests, which could provide the research community with instruments for creating arguments that could convince policy-makers to focus on important aspects of mathematics education that would otherwise not be easy to test.

Deduction and induction

The question pursued here has been one of *deductive* analysis: can a neural network use a specified way of coding data for analyzing new data? Another task of qualitative mathematics education would be the *inductive* generation of categories from data. In the AI field of unsupervised learning, neural networks are employed for discovering new patterns in data without human input. Results sometimes correspond to patterns recognizable by humans, and sometimes come as a surprise. Unsupervised learning thus holds a promise for the data-led generation of new descriptive knowledge, although owing to the black box character, any such knowledge would need to be interpreted very carefully.

Languages and big data. Such considerations, however, remain in the distant future. Much more research would be needed to investigate the feasibility of these hopes. This poses a significant challenge. The quality of a neural network is largely determined by the data used to train it. Data sets for natural language processing typically are quite large, comprising at least several tens of thousands or even up to millions and billions of samples. The data have to be labeled manually and consistently. Overall, data engineering represents a large area of work for AI. Additionally, language becomes a problem. Most work in AI is done using data in English language, and some crucial techniques are currently not available in many languages. Neural networks would need to be re-trained for each language. Simply translating available data into English would not present a solution, as language nuances important for mathematics education would likely get lost in the process. Additionally, this could lead to language-specific phenomena being discarded, which in turn could lead to a (further?) colonization of mathematics education research discourse by the English language. Data sets would need to be constructed for every language – which would be too large a task to handle for many research communities.

More questions than answers

The hopes for AI-based mathematics education research outlined above are high. But before their promise can be investigated in a methodically controlled way, much work of data engineering and theory building needs to be done. Too many theoretical questions are still open for the research community to adopt an AI-based approach. Neural networks seem capable of creating predictive knowledge by observing patterns, but can they also be used for establishing descriptive, prescriptive or explanative knowledge (Prediger, 2019)? Can researchers use them to generate explanations instead of only observations? Can they be used to generating new knowledge, or do they always reflect the old, either by the data used in training or through their constructor? Is the context-dependency of neural networks a handicap for research or is it actually an advantage? What is more important for research: definitions that satisfy theoretical rigor or black boxes that enable fast applicability? Is it ethical to use tools that are not fully understood? What methods, best practices, and methodological rules are needed to avoid the pitfalls of interpreting black boxes?

Although AI methods are state of the art regarding technological progress, they are freely available and can be applied even with very little programming experience. They do indeed provide almost ready-to-use tools. It remains the task of the research community to evaluate the use of these tools for mathematics education research.

References

Büscher, C. (2018). *Mathematical Literacy on Statistical Measures: A Design Research Study*. Wiesbaden: Springer. https://doi.org/10.1007/978-3-658-23069-2

Büscher, C. (2019). Conceptual learning opportunities in teachers' differentiated task designs for inclusive mathematics education. Paper presented at the *Eleventh Congress of the European Society for Research in Mathematics Education (CERME11)* in Utrecht, NL.

Gurevych, I., Meyer, C. M., Binnig, C., Fürnkranz, J., Kersting, K., Roth, S., & Simpson, E. (2018). Interactive Data Analytics for the Humanities. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing* (pp. 527–549). Cham: Springer. https://doi.org/10.1007/978-3-319-77113-7_41

Hiebert, J., & Grouws, D. A. (2007). The effects of classroom mathematics teaching on students' learning. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics* (pp. 371–404). Charlotte, NC: Information Age.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical Knowledge for Teaching and the Mathematical Quality of Instruction: An Exploratory Study. *Cognition and Instruction*, *26*(4), 430–511. https://doi.org/10.1080/07370000802177235

Kersting, N. B., Sherin, B. L., & Stigler, J. W. (2014). Automated Scoring of Teachers' Open-Ended Responses to Video Prompts: Bringing the Classroom-Video-Analysis Assessment to Scale. *Educational and Psychological Measurement*, *74*(6), 950–974. https://doi.org/10.1177/0013164414521634

Prediger, S. (2019). Theorizing in Design Research: Methodological reflections on developing and connecting theory elements for language-responsive mathematics classrooms. Avances de Investigación en Educación Matemática, 15, 5-27.

Prediger, S., Kuhl, J., Büscher, C., & Buró, S. (in press). Mathematik inklusiv lehren lernen: Entwicklung eines forschungsbasierten interdisziplinären Fortbildungskonzepts. *Journal für Psychologie*.

Rich, E., Knight, K., & Nair, S. B. (2009). *Artificial intelligence*. New Delhi: Tata McGraw-Hill.

Russell, S. J., & Norvig, P. (2010). *Artificial intelligence: A modern approach*. New Jersey: Pearson.

Wittgenstein, L. (1953/2003). *Philosophische Untersuchungen*. Frankfurt am Main: Suhrkamp.