Generalizing Expert Misconception Diagnoses Through Common Wrong Answer Embedding

John Kolb
UC Berkeley
jkolb@berkeley.edu

Scott Farrar*
Khan Academy / Independent
scottfarrar@gmail.com

Zachary A. Pardos
UC Berkeley
pardos@berkeley.edu

ABSTRACT

Misconceptions have been an important area of study in STEM education towards improving our understanding of learners' construction of knowledge. The advent of large-scale tutoring systems has given rise to an abundance of data in the form of learner question-answer logs in which signatures of misconceptions can be mined. In this work, we explore the extent to which collected expert misconception diagnoses can be generalized to held-out questions to add misconception semantics. We attempt this generalization by way of a question-answer neural embedding trained on chronological sequences of learner answers. As part of our study, we collect natural language misconception diagnoses from math educators for a sampling of student answers to questions within four topics on Khan Academy. Drawing inspiration from machine translation, we use a multinomial logistic regression model to explore how well the expert misconception semantics, in the form of bag-of-words vectors, can be mapped onto the learned embedding space and interpolated. We evaluate the ability of the space to generalize expert diagnoses using three levels of cross-fold validation in which we measure the recall of predicted natural language diagnoses across rater, topics, and questions. We find that the embedding provides generalization performance substantially beyond baseline approaches.

1. INTRODUCTION

The notion of mapping out abstract spaces of student learning and development has been around for ages, with Zone of Proximal Development [23] serving as a canonical example of defining the area of topics a student could learn with help from peers and the topics beyond. Work in Educational Data Mining has explored mapping out learning spaces taking the form of tree structures [4] or concept nodes in a directed graph [11], often used to represent prerequisite relationships. Other work has mapped out progress points within a course and their relationship to classical psychometric measures of ability [1]. In this work, we build on the idea of conceiving a space of learning as an embedding, or set of continuous vectors, with parts of the space indicative of different states of understanding and misconception [14]. We learn this embedding from sequences of millions of answers to exercises from a popular STEM tutoring system, then recruit qualified experts to diagnose a sampling of common wrong answers, providing natural language semantics to associate with question answers at their respective locations in the embedding. To test if the embedding generalizes these short form diagnoses, we use linear interpolation of the learned vector space to predict the words used in held-out diagnoses, holding out by expert, problem type, and question in cross-validation experiments. Successful predictive generalization in this task has implications for surfacing automatically generated misconception hypotheses to both teachers and computer tutors.

*At Khan Academy 2016-2017

John Kolb, Scott Farrar and Zachary A. Pardos "Generalizing Expert Misconception Diagnoses Through Common Wrong Answer Embedding" In: *The 12th International Conference on Educational Data Mining*, Michel Desmarais, Collin F. Lynch, Agathe Merceron, & Roger Nkambou (eds.) 2019, pp. 342 - 347

2. RELATED WORK

The theory of mathematical misconceptions described by Piaget [16], and considered by Smith, diSessa, and Roschelle [19] is one of continually developing partial understandings. Analysis of learner responses, rather than only correctness, may reveal aspects of their understandings. In the age of big data and computation, several modern approaches have brought different perspectives to the analysis of misconceptions. Feldman et al. [5] generated plausible production rules that could have produced the common wrong answers observed in student responses to addition questions in 11 elementary schools. In the vein of KC model or Q-matrix improvement [22], Liu, Patel, & Koedinger [6] explored adding KCs symbolizing buggy production rules to problem steps whose correct answer could be arrived at in spite of applying the buggy production. They found that the inclusion of this item-level misconception tagging improved the overall fit of their AFM model and the validity of the learned individual student parameters. Most complimentary to our work is the work of Michalenko, Lan, & Baraniuk [8], who did not study misconceptions in common wrong answers, but rather misconceptions found in the text of long open response text, using skip-grams and other embedding methods. Their approach is complementary to ours in that it cannot be applied to short, numeric answers in isolation. Inversely, our approach, which extends the embedding context across questions, is driven by questions that generate common wrong answers across students, which would exclude direct applicability to long answer response text.

2.1 Buggy Rules

In the cognitive theories underlying the design of intelligent tutoring systems [2], there are rules that produce correct and consistent answers, and efforts have been made towards cataloging collections of buggy rules that could instead produce incorrect answers. These buggy rules could represent misconceptions that students often have during the learning process [3]. This large collection of buggy rules is often referred to as a bug catalog [20]. As a student moves through a problem set, the bug catalog enables tutoring systems to tag, track, and respond to a path of answers the student provides.

Past research efforts to classify these buggy rules also include the manual labeling of misconceptions by experts [7], the exploration of cluster relationships between the wrong answers [15], and approaches that take into account the frequency of student misconceptions [21]. These efforts lay the foundation for automated approaches which utilize these buggy rules to generate targeted guidance messages specific to each incorrect answer [18].

2.2 Use of Skip-grams

Skip-gram models were originally applied to the embedding of words based on a large corpus of text (e.g. Wikipedia or a large archive of news articles). Once trained, the representational (hidden) layer of these models was shown to encode distributed concepts in the form of syntactic (e.g., bee is to bees as goose is to geese) as well as semantic relationships (e.g., Einstinen is to scientist as Picasso is to painter) [10]. While conventionally applied to language in its debut, skip-grams have been applied to non-linguistic data from education. University courses were embedded from sequences of enrollments [13] to find course similarities outside of what could be inferred from catalog descriptions. Questions within the ASSISTments tutoring platform were embedded based on sequences in which problems were answered in order to predict the skill of untagged questions [12]. Skip-grams and other embedding models have been applied to standard natural language in educational contexts, such as the learning of vector representations of open response text and correlating vector representations with the presence or absence of hand coded misconceptions[8].

3. TUTOR DATA SET

Our dataset of anonymized student answer logs comes from Khan Academy, an online STEM tutoring platform. As described in our previous work [14], Khan Academy categorizes student responses by *exercise*, a broad skill similar to those seen in ASSISTments Skill Builder sets; by *problem type*, a problem template; and finally by *seed*, one of two hundred values per problem type which uniquely identifies a template instantiation. Each log entry also contains an anonymous user ID and timestamp, which we use to group and chronologically sort student answers for model training.

We used the same exercise selection process as in [14] to narrow our focus to exercises with sufficient data and concerning topics that would likely surface interesting misconceptions for educators to analyze and describe. This involved consulting a subject matter expert in mathematical education [17] and verifying the correctness of the log entries by forming a sample set of questions and manually accessing their respective web pages on Khan Academy. At the conclusion of this filtering process, we identified four suitable exercises to use in our experiments:

1. "Surface Areas" (SA)
2. "Slope from an equation in slope intercept form" (SESI)
3. "Area of quadrilaterals and polygons" (AQP)
4. "Adding and subtracting fractions" (ASF)

Table 1 shows statistics for each exercise.

	SA	SESI	AQP	ASF
Problem Types	6	2	2	7
Seeds	38	20	50	40
Students	105,659	33,603	58,239	179,263
Unique Incorrect Answers	55,126	6,912	17,998	46,516
Total Incorrect Answers	619,045	112,390	298,356	873,916

Table 1: Descriptive statistics of exercises used to train the skip-gram models.

A second dataset was collected as part of this study, which consisted of natural language diagnoses of common wrong answers from our chosen exercises. These diagnoses were written by mathematics educators, with each diagnosis explaining the misconception that was potentially responsible for the incorrect answer. We collected misconception diagnosis labels using an online survey platform.[1] We describe the collection of these data in Section 4.2.

4. METHODOLOGY

In this section, we describe the techniques employed to complete three primary methodological tasks:

1. Generate learned question answer embeddings from student answer logs
2. Generate bag-of-words representations of the semantic data contained in educator diagnoses of the misconceptions associated with the incorrect student answers from (1.)
3. Compute a model that generalizes semantic diagnoses of wrong answers based on regression from the continuous vectors of (1.) to the semantic representations from (2.)

Figure 1 depicts the full data processing and machine learning pipeline that we implemented to complete these tasks, using both the answer event logs and the misconception diagnoses as inputs and outputting natural language diagnoses for held-out question answers.

4.1 Embedding Student Answers

As described in Section 2, machine learning models originally intended to model natural language have recently been applied to a number of other domains, including education. Motivated by the success of these efforts, we used a skip-gram neural network model to learn representations of student answers. A representation in our setting, or *embedding*, is a vector in a high-dimensional space that is learned by a skip-gram model. We use the same strategy as in [14] to encode each student answer in a token containing its *seed* and the

[1] `https://qualtrics.com`

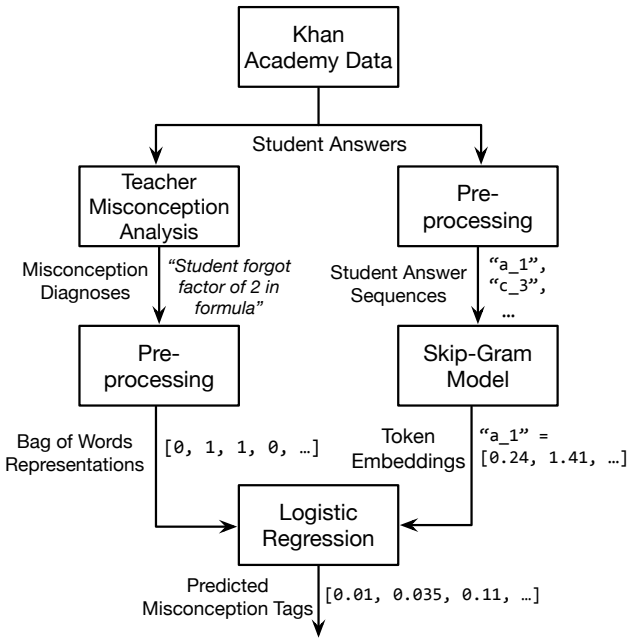

Figure 1: The pipeline used to model student answers, teacher diagnoses, and their correlation.

frequency rank of the student's response within that seed. For example, if a student were to answer a question generated from seed `x01b` with the most frequently occurring incorrect answer to that question, their answer would be represented by the token `x01b_1`.

A skip-gram model is a two-layer neural network (one hidden layer) that analyzes a corpus of token sequences to learn continuous vector representations for each of these tokens. Vectors are trained with the goal of predicting the context of each token. For example, `x01b_2` would have `s03c_4` in its context if students often provide incorrect responses to those questions in succession. The loss function (Eq. 1) for the training process, described in [10], seeks to optimize the log-likelihood of the tokens in context given a specific input token. S represents the set of input sequences for the model, each corresponding to a student's sequence of responses to a given exercise. c represents the window size, a hyperparameter of the model that specifies the width of a token's context when learning its representation, and T represents the number of tokens in sequence s.

$$C = -\sum_{s \in S} \frac{1}{T} \sum_{t=1}^{T} \sum_{\substack{-c \leq j \leq c \\ j \neq 0}} \log P(w_{t+j} | w_t) \qquad (1)$$

We use the negative sampling variant for training the skip-grams as introduced in [10], which replaces the final term of the form $\log P(w_O|w_I)$ in Equation 1 with

$$\log \sigma \left(v'^{T}_{w_O} v_{w_I} \right) + \sum_{i=1}^{k} \mathbb{E}_{w_i \sim P_n(w)} \left[\log \sigma \left(-v'^{T}_{w_i} v_{w_I} \right) \right] \qquad (2)$$

Above, σ represents the sigmoid function. Roughly, this formulation seeks to include the weights of k randomly chosen negative samples, i.e., tokens w_i that do *not* occur within the context of the target token w_O, in the backpropagation process. Unlike the original hierarchical softmax formulation, negative sampling has the advantage of only adjusting pairs of weights in the underlying network during backpropagation.

4.2 Collecting Teacher Diagnoses

We collected expert-generated semantic misconception diagnosis data through a questionnaire designed and run on the Qualtrics platform. Qualtrics recruited survey participants and compensated them on our behalf at a rate of $30 per participant. We had Qualtrics recruit participants who:

- Are working as a mathematics educator for students who are in grades 5–12 or undergraduates
- Have at least two years of prior teaching experience

The number of problem types and seeds within each exercise included in the survey is shown in Table 2. For each seed, we formed a batch of the five most frequently submitted incorrect answers to present to survey participants.

Exercise	# Prob. Types	# Seeds
Slope from an Equation in Slope Intercept Form	2	17
Adding and Subtracting Fractions	5	18
Surface Areas	6	36
Area of Quadrilaterals and Polygons	2	18

Table 2: Wrong answer exercises, problem types, and seeds for which expert diagnoses were sought

Each survey participant was provided with initial instructions, excerpted in Figure 2. Next, they were shown three randomly selected answer batches. For each batch, the survey respondent was presented with a screenshot of the original question as it appeared on Khan Academy, the text of the five incorrect student answers, and text boxes to write a brief misconception diagnosis for each answer. An example Khan Academy question and the associated diagnoses we collected are shown in Figure 3.

Respond with a general label-phrase that describes the most likely error or misconception related to the incorrect answer.

- Avoid references to specifics of the question (e.g., do not say "additive inverse is 4, not −4").
- Your label or phrase should be general enough such that it could potentially be applied to other incorrect answers. Therefore, you may duplicate labels and phrases as you see appropriate.
- Avoid abbreviations (e.g., use "y intercept" instead of "yint").

Example Responses
Question: Solve $3x - 4 = 20$
Student Answer: $5\frac{1}{3}$
Example Label-Phrase: opposite of additive inverse

Figure 2: An excerpt of the instructions presented to survey participants providing expert diagnoses

Alternatively, we could have asked experts to create misconception labels out of terms drawn from a fixed taxonomy,

$$-\frac{1}{4} - \left(-\frac{3}{5}\right) = \square$$

Answer	Misconception Diagnosis
17/20	Added $5+12$ instead of $5-12$.
$-17/20$	Added $5+12$ instead of $5-12$. And used incorrect sign.
$-7/20$	Has incorrect sign. Should be $+$.
2/5	Did not use common denominator.

Figure 3: A sample Khan Academy question and corresponding misconception labels

rather than to compose these labels from scratch and without explicit guidance. However, the terms in this taxonomy would inevitably reflect our own biases and assumptions and may prevent experts from accurately describing their observations. Instead, we allowed a broad vernacular, but also asked experts to review their labels at the end of the survey to encourage them to be consistent in their language.

We found that the quality of survey responses varied dramatically within our dataset and developed a procedure to identify and retain only misconception labels that were suitable for further analysis. We manually excluded all responses where an attempt at a label was clearly not present, such as "`idk`." Next, we retained diagnoses only from experts who wrote labels with an average length of 20 characters or more. This process left us with 570 unique diagnoses covering 14 of the 15 problem types and 64 of the 89 seeds.

4.3 Processing Teacher Diagnoses

After collecting expert misconception diagnoses through the survey platform, we performed data pre-processing to eventually represent each label in bag-of-words form. Many diagnoses contained references to specific numbers found in the instantiation of the question. We chose not to give every numerical quantity its own token but rather to replace each contiguous mathematical expression with the token `numN`, representing the N^{th} contiguous expression appearing in the diagnoses for each seed. Numbers used to describe general misconception rules, e.g. the factor of $1/2$ used in computing the area of a triangle, were hand-identified and allowed to be represented in original form. This helps to prevent our models from incorrectly identifying correlations that are coincidental (two question instances happen to use the same random quantity) rather than structural.

Next, we stripped punctuation, removed stopwords, and performed word stemming. Finally, we manually removed some of the most common tokens that we deemed uninformative and which could have resulted in trivially easy prediction due to their frequency, such as `student`, `tried`, and `used`. Each processed expert diagnosis is represented as a bag-of-words vector, where an element of the vector indicates the number of occurrences of a term from a global vocabulary. Where we had multiple expert labels available for a single incorrect student answer, we concatenated the two labels and constructed a bag-of-words representation of the result.

		Crossfold Type		
		Evaluator	**Prob Type**	**Seed**
Training	**Folds**	19	14	64
	Data Points	302	296	314
	Evaluators	18	19	19
	Exercises	4	4	4
	Prob Types	14	13	14
	Seeds	61	59	63
Test	**Data Points**	17	24	5
	Evaluators	1	3	1
	Exercises	2	1	1
	Prob Types	2	1	1
	Seeds	3	5	1

Table 3: Statistics for different cross validation schemes. Entries are rounded averages across folds.

4.4 Mapping Answer Vectors to Diagnoses

With both embeddings of student responses and expert-generated diagnoses in hand, we could explore the extent to which the continuous vector representation of an incorrect answer is related to a semantic description of the misconception underlying that answer. We trained a multinomial logistic regression model to calibrate this correspondence that uses a vector embedding of an incorrect student answer to predict the words in the expert's diagnosis of that answer. The regression takes as input an m-vector representing a student answer, where m is the dimensionality of the skip-gram embedding space (a hyperparameter of the model). The model produces as output an n-vector, where n is the size of the teacher misconception diagnosis vocabulary. Because of the regression's use of softmax, this n-vector forms a probability distribution across all terms used in the teacher diagnoses. The i^{th} element of the vector expresses the predicted probability that the i^{th} term of the diagnosis vocabulary applies to the student answer.

5. RESULTS

Here, we describe our results and methodology for evaluating the representations produced by a skip-gram model by using logistic regression and the expert-generated misconception diagnoses. We performed a search over the hyperparameters of the skip-gram algorithm and then compared the predictions generated by our machine learning pipeline to two baselines.

5.1 Skip-Gram Model Evaluation

Recall from Figure 1 that we use logistic regression to train a model identifying correlations between embeddings of student answers and semantic explanations of the underlying misconceptions responsible for incorrect answers. The model surfaces correlations by taking a vector representation as input and producing a probability distribution over the vocabulary of terms used by educators in their misconception diagnoses as output.

Using the semantic data collected from educators as ground truth, we evaluated the insights generated through logistic regression when using vectors produced by different skip-gram models as input. We performed a standard leave-one-out cross-validation (CV) procedure on the educator data. We then evaluated the quality of a model's predicted misconception tags for student answers in the remaining fold using recall at N, where the value of N for each prediction

is equal to the number of terms used in the original expert label for the relevant incorrect answer. This is defined as:

$$R = \frac{\left|\hat{T}_N \cap T\right|}{|T|} \quad (3)$$

where T is the set of terms contained in an educator's misconception diagnosis for an answer, $\hat{T}_N$ is the set of terms corresponding to the N largest entries in the probability distribution produced by the logistic regression when given an embedding of the answer as input, and $N = |T|$.

We performed three leave-one-out cross-validations using each of the following to determine the fold segmentation:

1. *Evaluator*: The ID of the educator who produced the misconception diagnosis.
2. *Problem Type*: The ID of the template used to generate a question.
3. *Seed*: The unique identifier of an instantiated question.

Descriptive statistics concerning the train and test splits for each scheme are summarized in Table 3.

5.2 Results of Hyperparameter Search

We trained over 750 skip-gram models using different combinations of hyperparameters and then ran each model through the cross-validation procedure described above. The hyperparameters we varied were:

1. *Vector Size*: The number of elements in the vector representations learned by the skip-gram model
2. *Window Size*: The width of each token's context, i.e., the number of surrounding tokens to consider in the loss function defined in Equation 1.
3. *Min Count*: The minimum number of times a token occurs in the training set to be included in the model.
4. *Training Epochs*

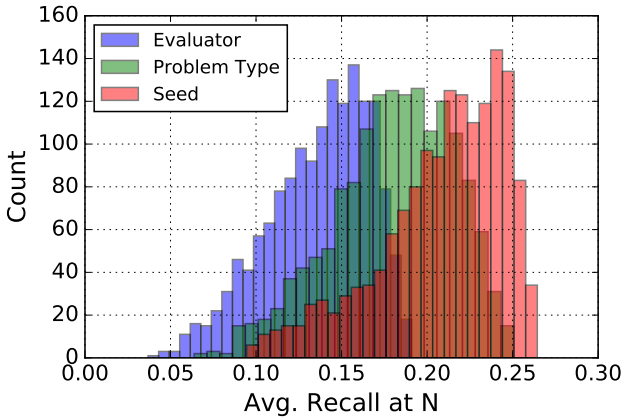

Figure 4: Distribution of average recall under the different cross-validation schemes.

Figure 4 shows the distribution of recall results achieved by all the models under each scheme. We also examined the distribution of hyperparameters among the ten models that achieved the highest average recall at N under each cross-validation type. We found that this metric was not sensitive to the hyperparameter values among the top ten models for all CV types. Within each CV type, all models produced scores within $0.0x$ of one another. Table 4 shows the hyperparameters that produced the best performing models, measured by average recall, for all CV types.

	Evaluator	Problem Type	Seed
Vector Size	60	100	100
Window Size	15	40	8
Min Count	10	15	5
Training Epochs	20	20	20

Table 4: The best skip-gram hyperparameter combinations under each cross validation scheme.

5.3 Diagnosis Generalization by Best Models

We compared the recall achieved by predicting the words in the diagnoses using the best skip-gram embeddings and logistic regression to the recall achieved by two baseline prediction schemes. For each incorrect student answer, all of the methods predict N terms, where N is the number of terms contained in the original expert diagnosis of the underlying misconception for that answer. This ensures we can fairly measure each prediction scheme by recall at N. The two baselines were:

1. *Random*: Generate a random sample of N terms from the vocabulary formed by the expert misconception diagnoses in the training set.
2. *Frequency*: Predict the N terms that appear most frequently in the diagnoses from the training set.

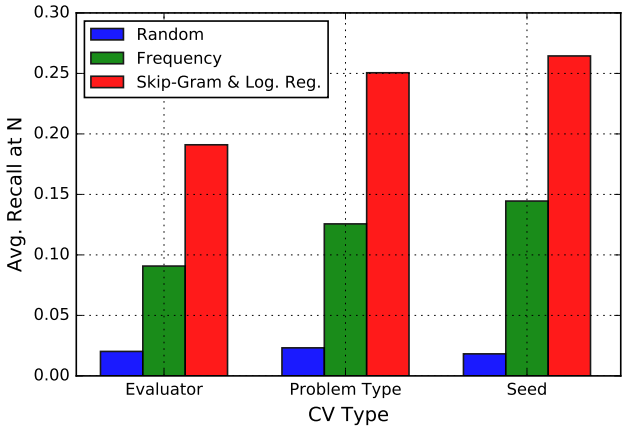

Figure 5: Average recall achieved by different prediction schemes for each cross-validation type.

The average recall at N achieved by the predictions generated through each baseline scheme, as well as that of our own approach, is shown in Figure 5. As expected, a frequency-based approach outperforms a random approach in all three cross-validation types. In addition, the embedding-based approach significantly outperforms the frequency-based approach in all three cases by nearly 100%. The results show that between 18% and 27% of words in held-out diagnoses were recovered. This improvement over baseline suggests a moderate correspondence between the regularities learned in the embedding and semantics used to describe misconceptions.

Recall increased with the size of the training set, with *Seed* having the largest training set and *Evaluator* having the smallest. Other factors may also contribute to these results. First, we chose Khan Academy exercises spanning a diverse selection of mathematical concepts, and the diagnoses for

misconceptions that arise in one domain (e.g., fractions) may use very different diagnosis terms than the terms used for misconceptions in another domain (e.g., surface area). Therefore, there are likely cases where the training set doesn't contain the proper terms to express the misconception diagnoses in the test set. Moreover, different educators used different taxonomies and terms when constructing their misconception diagnoses, which means a model may not be able to accurately predict the diagnoses provided by an educator that isn't well represented in the training data set, which appears to be the situation that arises in *Evaluator* cross-validation.

6. DISCUSSION

Should the 27% recall that we achieved in predicting the terms of held out misconception diagnoses be considered a good score? There are not prior results in this particular area with which to compare to a state of the art. However, this technique of linearly translating from one space (answer embedding) to another (diagnosis bag-of-words) is akin to machine translation from one language's embedding to another. Looking at the accuracy reported in the original linear machine translation paper [9], a translation accuracy of 10% was achieved between English and Vietnamese and 24% translated the other way. Therefore, we could consider 27% a comparable score to past NLP translation benchmarks and a performance level that may produce diagnoses that expert teachers could consider and potentially act on.

A limitation of our approach was that, as discussed in Section 4.2, our survey allowed experts to write open-ended misconception diagnoses which resulted in low frequency of some words and thus a more challenging downstream prediction task. A future study could restrict the terms available for use in expert labels or have them simultaneously negotiate a shared taxonomy. Finally, the student response sequences used as input for the skip-gram models were partitioned by Khan Academy exercise due to us wanting to focus on a limited number of topic areas. This may have lead to missing misconception signatures that manifest or generalize across exercises.

7. ACKNOWLEDGEMENTS

We thank Khan Academy for sharing anonymized exercise data. This work was supported, in part, by a grant from the National Science Foundation (#1547055).

8. REFERENCES

[1] R. Almond, I. Goldin, Y. Guo, and N. Wang. Vertical and stationary scales for progress maps. In *EDM*, 2014.

[2] J. R. Anderson. *Rules of the mind*. Psychology Press, 2014.

[3] J. S. Brown and K. VanLehn. Repair theory: A generative theory of bugs in procedural skills. *Cognitive science*, 4(4):379–426, 1980.

[4] M. Eagle and T. Barnes. Exploring differences in problem solving with data-driven approach maps. In *Educational Data Mining*, 2014.

[5] M. Q. Feldman, J. Y. Cho, M. Ong, S. Gulwani, Z. Popović, and E. Andersen. Automatic diagnosis of students' misconceptions in k-8 mathematics. In *CHI '18*, page 264. ACM, 2018.

[6] R. Liu, R. Patel, and K. R. Koedinger. Modeling common misconceptions in learning process data. In *Proceedings of the Sixth Intl. Conf. on Learning Analytics & Knowledge*, pages 369–377. ACM, 2016.

[7] T. S. McTavish and J. A. Larusson. Labeling mathematical errors to reveal cognitive states. In *European Conference on Technology Enhanced Learning*, pages 446–451. Springer, 2014.

[8] J. J. Michalenko, A. S. Lan, and R. G. Baraniuk. Data-mining textual responses to uncover misconception patterns. In *Proceedings of the 10th Conference on Educational Data Mining*, pages 208–2013, 2017.

[9] T. Mikolov, Q. V. Le, and I. Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013.

[10] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[11] A. Muehling. Concept landscapes-a new way of using concept maps. *Journal of Educational Data Mining*, 9(2):1–30, 2017.

[12] Z. A. Pardos and A. Dadu. Imputing kcs with representations of problem content and context. In *UMAP '17*, pages 148–155. ACM, 2017.

[13] Z. A. Pardos, Z. Fan, and W. Jiang. Connectionist recommendation in the wild: On the utility and scrutability of neural networks for personalized course guidance. *User Modeling and User-Adapted Interaction*, 2019.

[14] Z. A. Pardos, S. Farrar, J. Kolb, G. X. Peh, and J. H. Lee. Distributed Representation of Misconceptions. In J. Kay and R. Luckin, editors, *Proceedings of the 13th International Conference of the Learning Sciences (ICLS)*, pages 1791–1798, London, UK, 2018.

[15] R. Pelánek and J. Rihák. Properties and applications of wrong answers in online educational systems. In *EDM*, pages 466–471, 2016.

[16] J. Piaget. The child's concept of number, 1952.

[17] A. Schoenfeld. Personal Communication.

[18] D. Selent and N. Heffernan. Reducing student hint use by creating buggy messages from machine learned incorrect processes. In *Intl. Conf. on Intelligent Tutoring Systems*, pages 674–675. Springer, 2014.

[19] J. P. Smith, A. A. DiSessa, and J. Roschelle. Misconceptions reconceived: a constructivist analysis of knowledge in transition. *The journal of the learning sciences*, 3(2):115–163, 1994.

[20] W. L. J.-E. Soloway. Intention-based diagnosis of programming errors. In *Proceedings of the 5th National Conference on Artificial Intelligence, Austin, TX*, pages 162–168, 1984.

[21] M. Straatemeier et al. Math garden: A new educational and scientific instrument. *Education*, 57:1813–1824, 2014.

[22] K. K. Tatsuoka. A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational Statistics*, 10(1):55–73, 1985.

[23] L. S. Vygotsky. *Mind in society: The development of higher psychological processes*. Harvard university press, 1980. Original Manuscripts ca. 1930–1934.