Tracing Systematic Errors to Personalize Recommendations in Single Digit Multiplication and Beyond

Savi, O.A.; Deonovic, B.E.; Bolsinova, M.; van der Maas, H.L.J.; Maris, G.K.J.

DOI
10.5281/zenodo.5806832

Publication date
2021

Document Version
Final published version

Published in
Journal of Educational Data Mining

Tracing Systematic Errors to Personalize Recommendations in Single Digit Multiplication and Beyond

Alexander O. Savi
University of Amsterdam
O.A.Savi@gmail.com

Benjamin E. Deonovic
ACTNext
Benjamin.Deonovic@act.org

Maria Bolsinova
ACTNext
Maria.Bolsinova@act.org

Han L. J. van der Maas
University of Amsterdam
H.L.J.vanderMaas@uva.nl

Gunter K. J. Maris
ACTNext
Gunter.Maris@act.org

In learning, errors are ubiquitous and inevitable. As these errors may signal otherwise latent cognitive processes, tutors—and students alike—can greatly benefit from the information they provide. In this paper, we introduce and evaluate the Systematic Error Tracing (SET) model that identifies the possible causes of systematically observed errors in domains where items are susceptible to most or all causes and errors can be explained by multiple causes. We apply the model to single-digit multiplication, a domain that is very suitable for the model, is well-studied, and allows us to analyze over 25,000 error responses from 335 learners. The model, derived from the Ising model popular in physics, makes use of a bigraph that links errors to causes. The error responses were taken from Math Garden, a computerized adaptive practice environment for arithmetic that is widely used in the Netherlands. We discuss and evaluate various model configurations with respect to the ranking of recommendations and calibration of probability estimates. The results show that the SET model outranks a majority vote baseline model when more than a single recommendation is considered. Finally, we contrast the SET model to similar approaches and discuss limitations and implications.

Keywords: learning diagnosis, recommendation system, computerized adaptive practice, Ising model, ranking and calibration evaluation

1. Introduction

Students' errors provide a unique window into the mind, as error responses may reflect the cognitive processes—such as applied strategies—activated during problem solving. This fundamental understanding has spawned decades of research, from classifications of errors (Ben-Zeev, 1998; Straatemeier, 2014), and cognitive models aimed at explaining errors (Braithwaite et al., 2017;

Buwalda et al., 2016), to the diagnosis of observed errors (Taraghi et al., 2015; Taraghi et al., 2016). In this contribution to the field of errors in learning, we propose a model for the latter. The introduced Systematic Error Tracing model traces the latent causes of an individual student's manifest errors. It exploits a mapping of errors to causes—based on an analysis of errors and theories of multiplication in the developmental literature—that allows each error to be linked to multiple causes. Among other things, the approach benefits personalized recommendations.

1.1. Challenges in Error Diagnosis

Errors come in many shapes. One straightforward classification is the separation into unsystematic and systematic errors. An unsystematic error is usually termed a mistake or slip (Norman, 1981), that is, "the error that occurs when a person does an action that is not intended" (p. 1). This type of error may originate from sloppiness, carelessness, or inattentiveness. On the systematic end of the spectrum, one can distinguish so-called misconceptions, or rational errors (Ben-Zeev, 1995), that is, "students ... correctly following incorrect rules, rather than incorrectly following correct ones" (p. 342). Rational errors are sometimes described as bugs: incorrect perturbations of correct procedures (Brown and Burton, 1978). Ben-Zeev (1998) discusses various hypothesized origins of such errors, including the incorrect induction of examples (VanLehn, 1986). Finally, on top of misconceptions, there are decidedly different sources of systematic errors, including incorrect fact retrieval (McCloskey et al., 1991), heuristics (Reber et al., 2008), and biases (Shaki and Fischer, 2017), with the latter two gaining interest recently.

The ultimate promise of being able to diagnose the causes of error responses is the guidance it provides in adapting education to a student's individual needs. However, inferring the cause of an error—and ensuring that a student can benefit—poses serious challenges. The first challenge is to map all the possible causes of an error. Straatemeier (2014) elegantly illustrates this challenge with an example from arithmetic: $9 \times 9 = 18$. It is easily seen that the error can be caused by the use of the wrong operator; adding both operands rather than multiplying them. But, for all we know, the error might as well originate from the correct calculation, but the reversal of the tens and the unit digits in the answer. Such a tens-unit inversion exists, for instance in the Dutch language (van der Ven et al., 2017), where 81 is pronounced 'one-and-eighty'. Then, a third option is the *operator relevant* error, where 18 is the correct answer to the incorrect multiplication problem: 2×9 or 9×2. And to make things worse, in addition to these systematic errors, it could have simply been an unsystematic slip.

Graphically, this first challenge can be visualized by a so-called bipartite graph (or bigraph) that links causes to effects. The graph for the above problem is shown in Figure 1. This error-to-cause mapping links the selected errors to all of their possible causes (insofar that those causes are known). Figure 1 thus serves as a very minimal example, but nonetheless conveys the structure of a full error-to-cause mapping, which may include all single-digit multiplication errors and their known possible causes.

Additionally, the graph expresses the need to adapt education to the individual: although the error-to-cause mapping nicely summarizes the many different causes of observed errors across individuals, the actual causes may naturally differ from individual to individual, and from time to time. For example, one might retrieve the correct answer to an incorrect problem from memory, whereas another might have difficulty with transcoding. In many situations, such differences require different recommendations.

Importantly, the observed error $9 \times 9 = 18$ does not reveal the actual cause for one particular

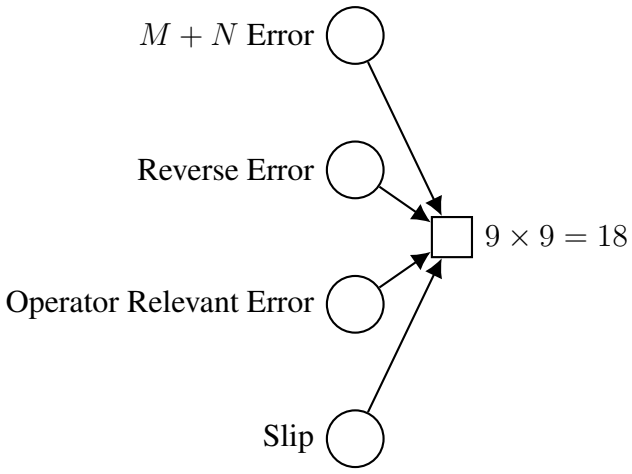

Figure 1: The bipartite graph for $9 \times 9 = 18$. The left column with circles depicts the latent causes and the right column with the square depicts the manifest effect. The arrows reflect the theory on what causes elicit which effect.

individual. Therefore, the second challenge, and the topic of this paper, is inferring the latent causes that drive the manifest errors of an individual. Again, this challenge can be displayed as a bipartite graph, shown in Figure 2. However, this time the error is represented by a black square, reflecting the fact that the error is observed, and the possible causes are represented by dashed circles, reflecting the fact that the actual cause for this particular individual is unknown.

Finally, the observed error responses pose a third challenge. Error responses are not ubiquitous, not necessarily consistent, and their causes may change over time. For the promise of adapting education to the individual, the inference of possible causes thus relies on limited data. This is not only a problem for inference, but also for evaluating the accuracy of a model. In this study, the latter problem is solved by using a response-intensive longitudinal data set, which allows us to investigate the robustness of the former.

1.2. Related Work

Since Corbett and Anderson's (1995) seminal introduction to knowledge tracing (KT), two and half decades of research have spawned advances in learner modeling, to a large extent summarized by Desmarais and Baker (2011) and Pelánek (2017). More recently, purely predictive approaches have been suggested (e.g. recurrent neural networks, Piech et al., 2015) and criticized for their limited explanatory power (Khajah et al., 2016). Here, we highlight two approaches.

One approach to the challenges of diagnosing errors is the use of cognitive diagnostic models (CDMs). CDMs are latent class models developed to identify the presence or absence of specific skills that are required to correctly answer a set of items. This is in contrast to traditional item response theory models, which measure ability on a unidimensional scale. Instead of measuring a single unidimensional rating for each person, CDMs maintain a profile $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ where $\alpha_k = 1$ if the kth skill has been mastered and $\alpha_k = 0$ if the kth skill is not mastered for $k = 1, \ldots, K$. A key construct in CDMs is the Q-matrix that specifies which skills are required by which items (Tatsuoka, 1983). Various different CDMs exist but differ in how they relate the latent class profile and the Q-matrix to the observed responses. Examples of CDMs include the DINA model (Haertel, 1989) and the DINO model (Templin and Henson, 2006).

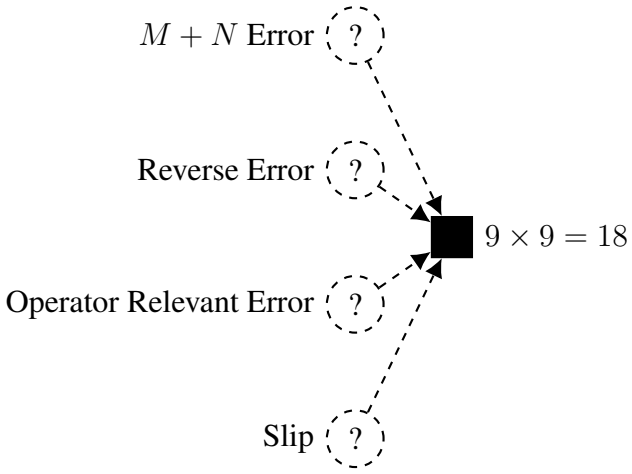

Figure 2: The problem of identifying the cause of an error for a particular individual, represented in the bipartite graph for $9 \times 9 = 18$. The left column with dashed circles depicts the possible latent causes and the filled square in the right column depicts the observed effect.

In recent literature, CDMs have been extended to identify misconceptions, in some cases in tandem with skills (Bradshaw and Templin, 2013; Kuo et al., 2016; Kuo et al., 2017). However, the limitation of CDMs for identifying misconceptions is in the use of the Q-matrix to relate misconceptions to items. Importantly, all models that have used CDMs to identify misconceptions map those misconceptions to specific items, severely limiting the kind of misconceptions that can be analyzed. From the previously discussed bigraph, it is seen that misconceptions are often not related to a particular *item*, but rather to a specific *error*. Again, consider the item 9×9; many causes can elicit an erroneous response to this item (in fact, all error categories in Table 1, which is discussed in the Methods section). On the other hand, if we know the specific error that was made was $9 \times 9 = 0$ we know a *zero* error or *m minus n* error occurred. To quantify this limitation, Figure 3 shows for all single-digit multiplication items, the number of considered error categories that may apply. It shows that all considered error categories can apply to the largest fraction of items: 50 items are susceptible to all 14 considered causes, 21 items are susceptible to 13 of the considered causes, and 10 items are susceptible to 12 of the considered causes.

In fact, this relationship between causes, errors, and items precludes the use of CDMs in all but the most trivial examples, at least for the type of data considered herein. The necessary and sufficient conditions for identifiability of parameters in CDMs have recently been derived (Gu and Xu, 2018), and when there exist causes that are tagged to all items these identifiability conditions are not met. Consequently, the estimation of parameters in such a setting would be suspect and lead to spurious results. Therefore, the fundamental unit of analysis in this work is the errors that are exhibited by particular items. Considering these more specific events allows us to analyze the sparser associations between causes and errors rather than the dense association between causes and items.

That is not to say that modeling error responses rather than item responses in a CDM is theoretically impossible. However, it would require a polytomous CDM to account for the multiple possible ways an item could generate an error (or correct response). Such models have been recently described in the literature (Chen and de la Torre, 2018). The current implementations assume the polytomous responses are ordered, which would not be an appropriate assumption

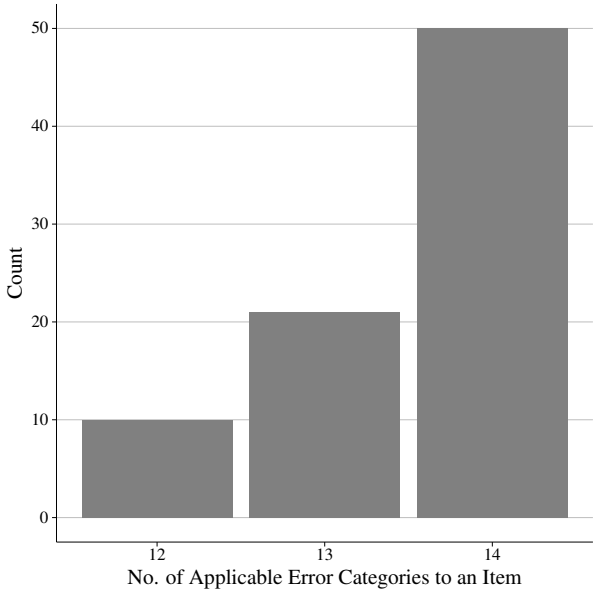

Figure 3: Histogram of the number of error categories that can apply to a single-digit multiplication item.

for different error responses to an item. To allow for a *nominal* polytomous response, CDMs would have to be developed which allow a separate Q-matrix to be specified for each response category.

On top of that, another issue precludes CDMs from being an appropriate method for analyzing data described herein. CDMs are fundamentally models for cross-sectional data rather than longitudinal data. They assume that the latent attributes (whether they are misconceptions or skills) are fixed and do not vary over time. Although the approach utilized in this paper (applying the model in a sliding window of time) could be applied to CDMs as well, this is not possible with off-the-shelf software and may result in too few responses in a window frame to estimate the parameters of a CDM.

Contrary to CDMs, Bayesian Networks (BN; Pearl 1988; Vomlel 2004; Xu and Mostow 2011) are a popular and powerful approach that does allow one to model data that can be expressed as in the graphical model of Figure 1. BNs are actively used in different types of tutoring systems, including Cognitive Tutors (Conati et al., 2002) and Constraint-Based Tutors (Mitrovic, 2011). Since any BN can be converted into a Markov random field through graph moralization (Cowell et al., 1999), we do not consider the BN approach different from the one described herein.

1.3. Systematic Error Tracing

In this paper, we introduce and evaluate an approach that can be used to diagnose an individual's error responses: the Systematic Error Tracing (SET) model. The SET model can be seen as a knowledge-based recommendation system—or expert system—that exploits known relations between causes and errors as captured in a bigraph, in order to suggest the possible causes of the observed error responses. In essence, the model takes the proportion of connections between a cause and a set of observed errors as an indication of that cause's probability. Ultimately, this

ensures that the model is simple and intuitive, and allows for easy implementation in online learning systems.

The SET model derives its name from the fact that it traces and diagnoses *systematic errors*. As described in the second paragraph, systematic errors do not solely originate from what is generally viewed as a misconception: an error in the *understanding* of a concept or procedure. With the proposed model, any cause that systematically produces errors can be traced. An example is the systematic repetition or omission of a digit in the otherwise correct response (i.e., a *typo* in the terminology of Straatemeier, 2014).

This choice for systematic errors is deliberate. Rather than explaining exactly how and why errors occur—the objective of a cognitive model—the proposed model capitalizes on its diagnostic strength. It is not tied to a specific theoretical framework (such as rational errors, heuristics, or biases) and is unconstrained with regard to the included error categories, in order to allow a broad diagnosis of errors. The SET model is discussed in detail in the Methods section.

1.4. Application and Evaluation

In demonstrating the model, we apply it to the problem of single-digit multiplication. Although in principle it works in any domain, as long as clear errors can be identified and the possible causes can be mapped—that is, a bigraph like in Figure 1 can be created—not all domains lend themselves that easily. Multiplication serves as a great illustration for a couple of reasons. First, the very procedural nature of multiplication allows for the identification of clearly defined errors, which in turn has motivated scholars to identify a multitude of causes. Second, adding to this convenience, software algorithms can easily and automatically detect multiplication errors. Indeed, many online learning systems exist that provide multiplication education, and could thus readily benefit from the method. Finally, we focus on *single digit* multiplication for important reasons. Many educational methods, both online and offline, treat single digit multiplication as a separate domain, and it moreover helps us maintain a clear focus on the proposed model. Indeed, notably fewer studies have targeted multi-digit multiplication errors and the number of possible causes in single-digit multiplication already introduces quite some complexity.

To illustrate and evaluate the model, we use data from Math Garden, a computer-adaptive practice environment primarily aimed at Dutch primary school children. At the time of data collection, their single digit multiplication domain has generated over 25 million responses from over 170 thousand different primary school children, spanning 3 years, and with over 5 million errors made. This learning environment originates from an innovation that enables the adaptive assignment of problems to students, based on on-the-fly updated general measures of ability and difficulty, by means of an adapted Elo rating system (Klinkenberg et al., 2011; Brinkhuis et al., 2018). The required error-to-cause mapping was adapted from an existing error classification and is introduced in the Methods section.

2. Methods

In this section, we provide the employed error classification, describe the obtained data, introduce the model that we use to trace the causes of the observed errors, and discuss the procedure we use to evaluate the model. Note that we use the terms causes and error categories to describe the common cause or shared category of particular error responses; we use the terms observed

Table 1: The error categories considered in this study. Adapted from Straatemeier (2014) for single-digit multiplication.

Cause	Description
Operator Relevant	Answer to a different single digit multiplication problem, from the tables 2 to 9
Operand Related	Answer to a problem with one matching operand, and one operand that is 1 or 2 units smaller/larger (except when operand becomes zero or negative)
Double Half	Double or half the correct answer
Different Unit	Answer has incorrect unit
Miss 1	Correct answer plus/minus 1
Miss 10	Correct answer plus/minus 10
Miss 100	Correct answer plus/minus 100
Miss Power	Correct answer with the decimal point misplaced by 1 to 5 positions (i.e., correct answer multiplied/divided by 10 to the power of 1 to 5)
M Div N	The first operand divided by the second operand
M Minus N	The first operand minus the second operand
M Plus N	The first operand plus the second operand
Typo	Correct answer with the repetition or omission of a digit (omission only when correct answer has 2 digits or more)
Reverse	The digits of the correct answer are reversed (only for problems with a solution that consists of 2 digits)
Zero	0 is (incorrectly) provided as the answer

errors, error responses, and effects to describe students' actual error responses; and we use the terms students, children, and users to describe the users that provided the error responses.

2.1. Mapping

2.1.1. Error classification

As discussed in the introduction, theory links the errors to their possible causes. This error-to-cause mapping is captured in a bigraph, of which Figure 1 gives an illustrative example. For this bigraph, we used the error classification proposed by Straatemeier (2014). Conveniently, their classification is based on the rich Math Garden data, which enabled them to identify a great many causes. As there is no principled method to create such a classification, we took a pragmatic approach in adapting their classification to suit single-digit multiplication. Also, we only considered the categories for which Straatemeier (2014) identified *systematic* responses. Table 1 provides the selection of error categories and their definitions.

Some causes link to more errors than others. Table 2 gives the number of errors that causes can explain, divided by the total number of unique errors that all causes combined can explain. Importantly, these proportions do not sum to one, for the significant overlap between causes as discussed in the introduction.

Table 2: The number of errors a cause can explain, proportional to the number of errors that all considered causes can explain. Proportions do not sum to 1 due to the considerable overlap.

Cause	Proportion
operator_relevant	0.618
miss_power	0.176
different_unit	0.159
operand_related	0.119
typo	0.056
double_half	0.035
miss_1	0.035
miss_10	0.035
miss_100	0.035
m_minus_n	0.018
zero	0.018
m_plus_n	0.017
m_div_n	0.016
reverse	0.011

2.2. DATA

2.2.1. Math Garden

Math Garden hosts a variety of domains, primarily related to arithmetic, that can be practiced in isolation. One such domain is the *multiplication table* domain, which provides the data for this study. The 22 other domains include problems ranging from word problems to logical reasoning tasks. Importantly, the previously mentioned adaptive algorithm matches students to problems. Students can be viewed as competing with the problems in a domain, and the outcome—both in terms of speed and accuracy (Maris and van der Maas, 2012)—feeds into the adapted Elo algorithm, to continuously update student ability estimates and problem difficulty estimates.

In the multiplication table domain, students practice the multiplication tables of one to ten. For 5, 10, 15, or 20 seconds, a problem is presented, and during this time the student can provide a solution by means of a visualized numeric keypad. For each second that there is time left to solve the problem, the student receives a virtual coin for a correct response and loses a virtual coin for an incorrect response. This scoring rule represents the implemented speed-accuracy trade-off and is visualized by coins disappearing from the screen. Too difficult problems can be skipped by using a question mark button, which is without consequence.

Importantly, Math Garden is primarily used in natural learning settings. Both schools and families can buy subscriptions, and the system is used in and outside of the formal school setting. This property, along with the sheer amount of problems solved, creates a unique data set with very diverse error responses (previously analyzed by Straatemeier, 2014).

2.2.2. Selection criteria

These data properties require a careful selection procedure. First, we restricted selection to a three months period (January to March 2017), students that allowed scientific research based on their responses, and students in Dutch grades 3 to 8 (comparable to grades 1 to 6 in the US, and

Table 3: Numbers of students and average response frequencies per grade.

Grade	No. of Students	Avg. No. of Error Responses	Avg. No. of Correct Responses
3	16	74.88	450.25
4	42	62.83	443.52
5	53	87.09	576.85
6	64	77.31	715.25
7	84	73.92	671.62
8	92	78.92	982.71

approximately age 6 to 12). This is the same grade and age range that was used to create the classification and is justified by huge individual differences in ability estimates, with significant proportions of students scoring below the mean ability in the one or two grades below them, or scoring above the mean ability in the one or two grades above them (Straatemeier, 2014, p. 13).

Then, three factors guided the remaining procedure: the domain to which the model was applied, the model itself, and the evaluation procedure. To begin with, we applied the model to single-digit multiplication problems, and thus omitted all responses to the multiplication table of ten. Then, the model only takes error responses into account, and we thus omitted correct responses, question mark responses, and non-responses (time-outs). Also, we omitted error responses that could not have been elicited by any of the causes in the used classification and instantaneous errors (< 1 second).

Finally, the in vivo nature of the learning environment demanded a somewhat arbitrary final selection. From the 26,106 remaining students, we only selected the active and committed students that permitted a proper evaluation of the model. We selected students that practiced a minimum number of problems (≥ 80 problems, 4,511 active students) and made a minimum number of errors (≥ 40 errors, 348 students). The windowed evaluation (see Evaluation section) demands that a minimum number of errors is observed. Given this selection, we allowed a maximum proportion of error responses ($\leq 50\%$ errors, 346 committed students). Importantly, these final steps are highly pragmatic and may limit the generalization of the results to students that practice regularly (active students) and for whom the majority of responses are correct (committed students).

2.2.3. Error response distributions

Table 3 gives the average number of error responses per student and per school grade in the selected data. To put these numbers in perspective, the table also gives the average number of correct responses per student and per school grade.

Figures 4 and 5 show histograms for the number of error responses per student and per error category. Figure 4 shows the number of error responses per student, with a total of 346 students in the selected data. In the figure, note that students with less than 40 error responses were filtered out and that the distribution is highly skewed with a few students having provided many error responses.

Figure 5 shows the number of error responses per error category. This distribution allows for two different representations. In the top panel, the total number of error responses across error categories is shown. This includes duplicate errors (across and within students) and thus shows the breadth of the error categories across all observed errors. In the bottom panel, the number

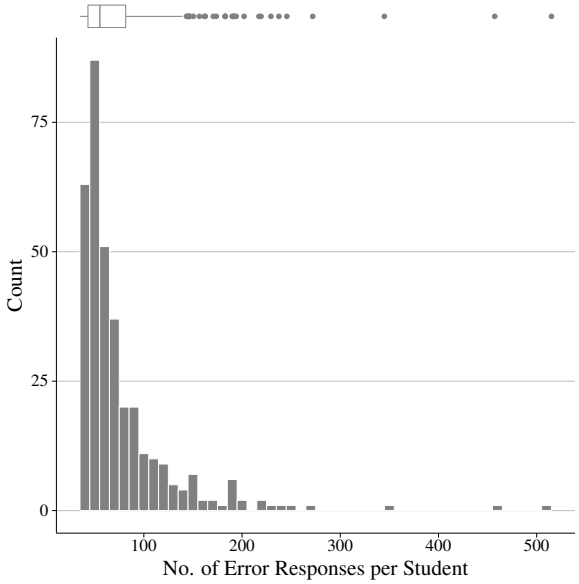

Figure 4: Distribution of the number of error responses per student. Bin width $= 10$. At the top of the figure, the histogram is represented as a box-and-whisker plot.

of unique error responses is shown. Here, all duplicate errors have been removed, thus showing the breadth of the error categories for unique errors.

Finally, in Figure 6 we use the UpSet visualization technique (Conway et al., 2017) to show the non-empty intersections for the five most dominant error categories. Key to the challenges addressed in this paper is the fact that a single error can have multiple origins. As a result, error categories overlap. With few categories, Venn or Euler diagrams can conveniently map all intersections. With many categories, however, relevant selections of the intersections have to be made. The UpSet diagram shows the non-empty intersections.

2.3. MODEL

2.3.1. Systematic Error Tracing

An intuitive understanding of the SET model's primary mechanism is easily obtained. Given a subset of observed errors, one simple method to calculate the probability of a cause is to determine the number of errors associated with the cause of interest, proportional to the total number of associations between causes and errors. Figure 7 exemplifies this method, showing that it comprises of no more than computing the proportion of edges for each of the causes in the observed bigraph. The model we introduce hereafter shares the idea that the relative number of links between a cause and the observed errors provides a proxy for the plausibility of the cause. Importantly, assigning the obtained probabilities to the considered causes is in accordance with Luce's choice axiom (Luce, 2005). This axiom states that the probability of selecting one item over another from a pool, should not be affected by which items are present in the pool. Such probabilities are said to have independence from irrelevant alternatives.

Now, let $G = \{C, E, A\}$ be a bipartite graph where C is the set of nodes related to causes, E is the set of nodes related to errors, and A is the set of (weighted) edges relating the causes and errors. We can represent A in matrix form such that the i, j element, a_{ij}, is the weighted edge

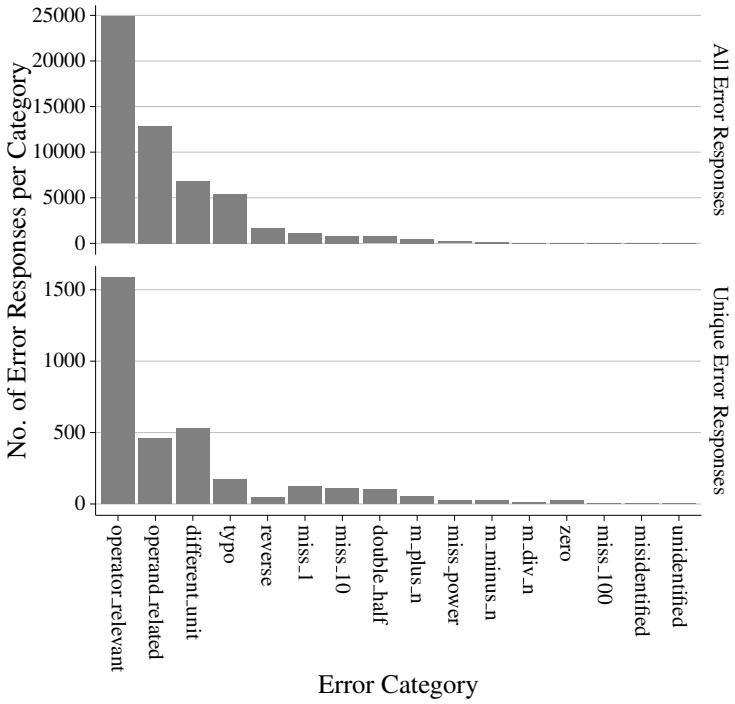

Figure 5: Distributions of error responses across error categories. The top panel shows the total number of observed error responses (y-axis) that belong to a certain error category (x-axis). The bottom panel shows the number of uniquely observed error responses that belong to a certain error category.

from cause i to error j for $i = 1, \ldots, n$ and $j = 1, \ldots, m$. A node $e \in E$ can either be $e = 1$ or $e = 0$ to indicate whether the error has been observed or not observed. A node $c \in C$ can likewise be either $c = 1$ or $c = 0$ to indicate the presence or absence of the respective cause.

We model the joint distribution of causes and errors as a type of Ising model (Ising, 1925). The Ising model—also known as the quadratic exponential binary distribution (Cox and Wermuth, 1994)—is a simple model for jointly modeling the distribution of a set of dichotomous variables. It was originally formulated to model ferromagnetism in physics. The standard Ising model is for variables that are coded with ± 1 whereas we use $1/0$ encoded variables. The joint distribution can be expressed as

$$p(\mathbf{x} = (\mathbf{c}, \mathbf{e}) | \boldsymbol{\mu}, \beta) = \frac{1}{Z} \exp(\beta \mathbf{x}^\intercal \boldsymbol{\mu} + \beta \mathbf{x}^\intercal \boldsymbol{\Sigma} \mathbf{x}) \tag{1}$$

where $\mathbf{x}^\intercal = \begin{pmatrix} \mathbf{c}^\intercal & \mathbf{e}^\intercal \end{pmatrix}$, $\boldsymbol{\mu}$ are the parameters associated with causes and errors (the external magnetic field in the Ising literature), Z is the normalizing factor, and $\boldsymbol{\Sigma}$ is the interaction effect matrix where the i, j element of $\boldsymbol{\Sigma}$, σ_{ij}, corresponds to the interaction strength between x_i and x_j.

Because G is bipartite, $\boldsymbol{\Sigma}$ is of the form

$$\boldsymbol{\Sigma} = \begin{pmatrix} \mathbf{0} & A \\ A^\intercal & \mathbf{0} \end{pmatrix}$$

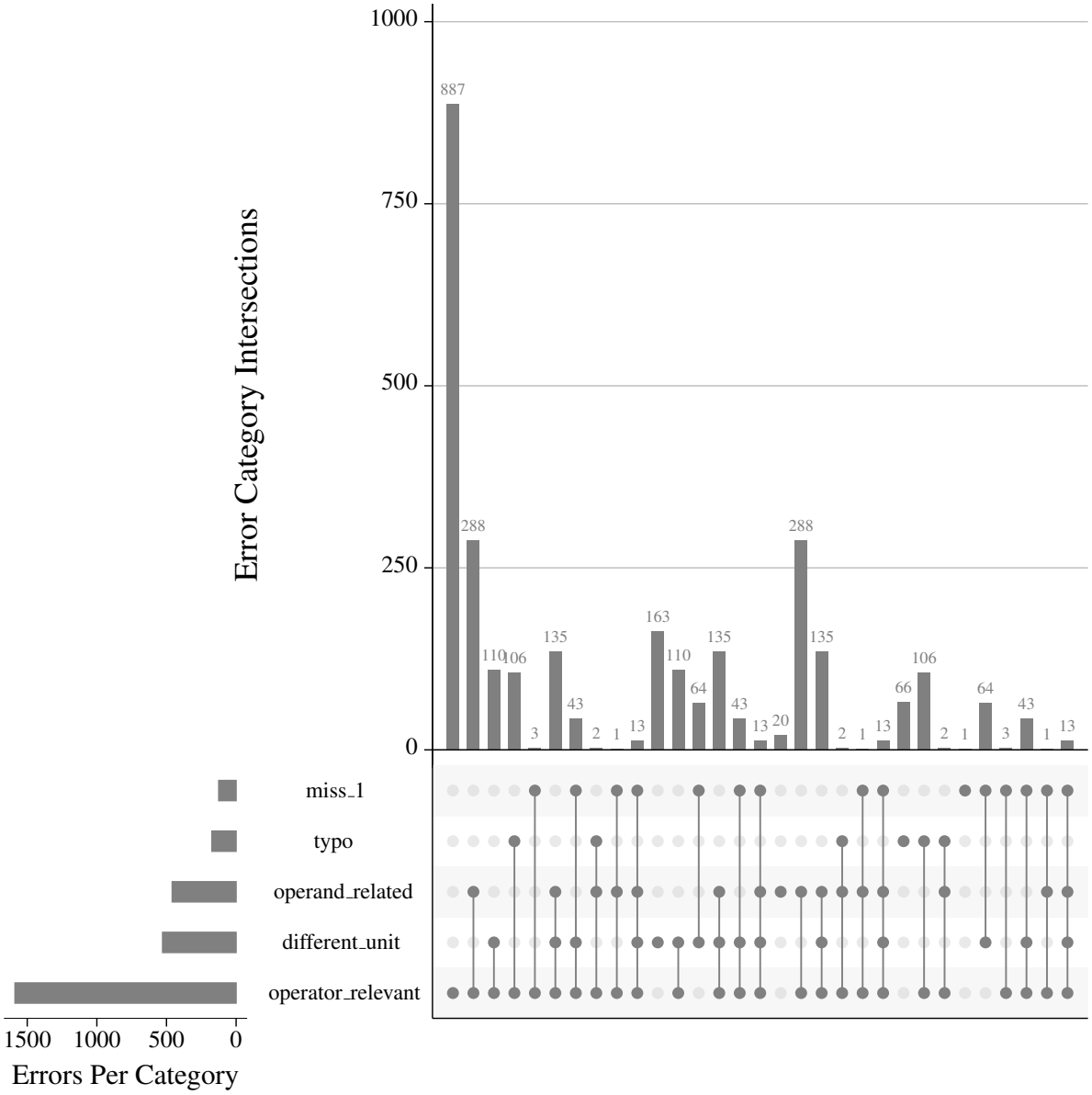

Figure 6: Intersections of the error categories. Bottom: the five most dominant error categories, with the number of unique errors per category (left), and various non-empty intersections of those categories (right). Top: frequency distribution of the errors in each non-empty intersection. The frequency distribution shows (from left to right) the largest error category, its intersections with one other category ordered by size, its intersections with two other categories ordered by size, and so on.

so we have

$$p(\mathbf{x} = (\mathbf{c}, \mathbf{e})|\boldsymbol{\mu}, \beta) = \frac{1}{Z} \exp\left(\beta \sum_{i=1}^{n} \mu_i c_i + \beta \sum_{j=1}^{m} \mu_{n+j} e_j + 2\beta \sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij} c_i e_j\right) \tag{2}$$

Ultimately, we want to be able to give recommendations about which cause should be targeted in an intervention, under the assumption that there is a single cause present. Thus, provided that we are interested in the probability that a particular cause is present given that exactly one cause is present, and given that we observed all considered errors in the data, we define

$$p_i = p(c_i = 1| \textstyle\sum_{i=1}^{n} c_i = 1, \mathbf{e} = \mathbf{1}, \boldsymbol{\mu}, \beta) \tag{3}$$

$$= \frac{p(c_i = 1, \sum_{l=1, l\neq i}^{n} c_l = 0, \mathbf{e} = \mathbf{1}|\boldsymbol{\mu}, \beta)}{\sum_{k=1}^{n} p(c_k = 1, \sum_{l=1, l\neq k}^{n} c_l = 0, \mathbf{e} = \mathbf{1}|\boldsymbol{\mu}, \beta)} \tag{4}$$

$$= \frac{\exp\left(\beta\mu_i + \beta \sum_{j=1}^{m} \mu_{n+j} + 2\beta \sum_{j=1}^{m} a_{ij}\right)}{\sum_{k=1}^{n} \exp\left(\beta\mu_k + \beta \sum_{j=1}^{m} \mu_{n+j} + 2\beta \sum_{j=1}^{m} a_{kj}\right)} \tag{5}$$

$$= \frac{\exp\left(\beta\mu_i + 2\beta \sum_{j=1}^{m} a_{ij}\right)}{\sum_{k=1}^{n} \exp\left(\beta\mu_k + 2\beta \sum_{j=1}^{m} a_{kj}\right)} \tag{6}$$

Figure 7 provides an example. With $\mu_i = 0$ for $i = 1, \ldots, 4$, $\beta = 1/2$, and $a_{ij} = 1$ ($a_{ij} = 0$) for connected (disconnected) causes and errors, we have

$$p_1 = \frac{\exp(1)}{\exp(1) + \exp(1) + \exp(4) + \exp(3)} \qquad p_2 = \frac{\exp(1)}{\exp(1) + \exp(1) + \exp(4) + \exp(3)}$$

$$p_3 = \frac{\exp(4)}{\exp(1) + \exp(1) + \exp(4) + \exp(3)} \qquad p_4 = \frac{\exp(3)}{\exp(1) + \exp(1) + \exp(4) + \exp(3)}$$

2.3.2. Strengths and limitations

In addition to being intuitive, the model has several benefits. To begin with, the Ising model provides the simplest model of a network of pairwise correlations with binary nodes. It captures the desired correlations between causes and errors, without the need to impose parametric assumptions on the data. The Ising model is an exponential family model and is the simplest such model that preserves the first and second moment on a network of binary nodes. In this sense, it is the binary analog of the multivariate normal distribution. Second, it is very easy to use. Because the univariate conditional distributions have simple and closed forms, one may obtain samples from the joint distribution using Gibbs sampling. This allows one an easy and efficient procedure to calculate the necessary values for identifying the most likely causes. Third, the SET model neatly accounts for the fact that a slip could have caused the error, which we show in Appendix A. Slips are therefore not explicitly included in the error-to-cause mapping. Fourth, the model can easily be generalized to probability for a given number of causes. For example,

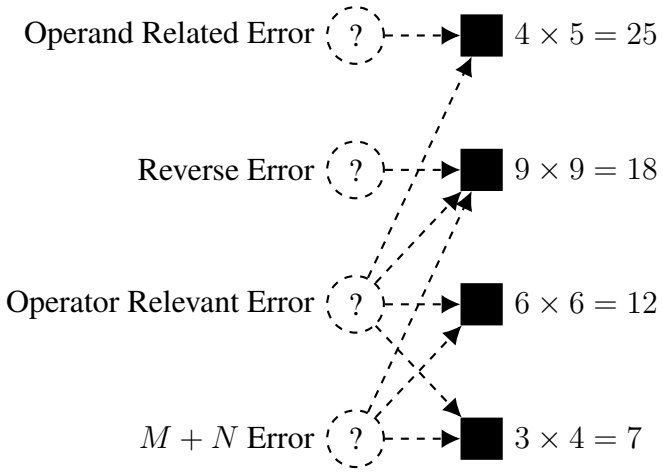

Figure 7: An example bigraph with 4 causes and 4 errors.

the probability that a particular pair of causes is present given that exactly two causes are present is defined as

$$p_{i_1,i_2} = p(\text{cause } i_1 \text{ and } i_2 \text{ are present}|\text{exactly two causes are present}) \tag{7}$$

$$= p(c_{i_1} = 1, c_{i_2} = 1| \textstyle\sum_{i=1}^{n} c_i = 2, \mathbf{e} = \mathbf{1}, \boldsymbol{\mu}, \beta) \tag{8}$$

$$= \frac{\exp(\mu_{i_1} + \mu_{i_2} + \sum_{j=1}^{m} a_{i_1,j} + \sum_{j=1}^{m} a_{i_2,j})}{\sum_{k_1=1}^{n-1} \sum_{k_2=k_2+1}^{n} \exp(\mu_{k_1} + \mu_{k_2} + \sum_{j=1}^{m} a_{k_1,j} + \sum_{j=1}^{m} a_{k_2,j})} \tag{9}$$

Another characteristic of the SET model is its bias to appeal to majority. This may occur when two or more causes are present. Figure 8 shows three causes and two observed errors. The two filled nodes are active for this particular student, each causing one of the errors. However, in this particular instance, the model would predict the inactive cause to cause the two observed errors, as it outweighs the other two causes in its proportion of edges. This dynamic illustrates a dominance of causes that are related to many different errors over causes that are related to a few errors. Having said that, in most cases this appeal to the majority makes perfect sense. Given that one can assume a single cause, the one cause that explains the most errors simply serves as the most parsimonious recommendation.

A second example of this appeal to majority is apparent in the misidentification of slips. A true slip may have a higher probability of being misidentified as a dominant cause—a cause that is related to many different error responses—than as a cause that could have for instance only produced a single error response. For example, only one slip response to the item 9×9 will be misidentified as an *m* + *n* error—the response 18—whereas evidently many more slip responses will be misidentified as an *operator relevant* error. Importantly, the misidentification of slips does not necessarily result in an appeal to majority, but assuming slips are truly random it does.

Contrary to models that use the item as the fundamental unit of analysis, including CDMs, the SET model is confined to items that elicit an erroneous response. As previously explained, the model is used for identifying systematic errors and is especially powerful in situations where the item itself provides no information. However, there is no fundamental limitation of using other sources of information—such as response times—in the bigraph. Also, the model is updated after each newly observed error response, even if the item and/or error response was pre-

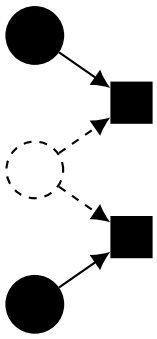

Figure 8: A bigraph with three causes (circles) and two observed errors (black squares). The black circles represent active causes, whereas the white circle represents an inactive cause. The SET model incorrectly predicts the inactive cause to elicit the two errors.

viously observed. Evidently, this is not a fundamental requirement either and may be relaxed. Finally, the SET model cannot predict a cause that is not part of the bigraph.

2.4. Evaluation

2.4.1. Baseline model

The previously described appeal to majority dynamic led us to compare the SET model to a baseline model that simply takes a majority vote. The majority vote (MV) model was created by determining the number of single-digit multiplication errors a cause can elicit proportional to the number of unique errors that the considered causes combined can elicit. Causes were then ordered by proportion. Table 2 gives these proportions for all considered causes.

2.4.2. General procedure

Table 4 summarizes the methods used to evaluate the SET model. In the following, we outline these methods in full detail. For evaluative purposes, we treated the SET model as a multi-label classification method. As such, model performance was determined by predicting the potential causes of a student's observed error responses from her or his preceding error responses. We used the following general procedure. For each individual, we first determined the *expected causes* by gathering a predetermined number of successive error responses (the *error window*) and determining the model's expected cause ranking from the responses in that window. This ranking was established by ordering the predictions of the model (estimated probabilities for the SET model and overall edge proportions for the MV baseline model). In the error window, we disregarded the passed time and possible correct responses in between successive error responses.

Second, we determined the *potential causes* by gathering the first error response following the evaluated error window and determining the possible causes for the selected error using the classification scheme discussed in the Mapping section. We moved the error window with a single error response at a time. For example, with an error window of size 15, we took the first 15 errors of a student to predict the 16^{th} error, we then took the second to 16^{th} error to predict the 17^{th} error, and so on. We evaluated five different window sizes, consisting of 1, 3, 7, 15, or 30 errors. Evidently, the window size cannot affect the performance of the MV baseline

Table 4: Summary of the evaluation methods.

Type	Element	Description
Models	MV	The Majority Vote baseline model. Predicted causes were ordered by the proportion of errors they cover. The proportions are shown in Table 2.
	SET	The Systematic Error Tracing model. Predicted causes were ordered by the SET model's expected probabilities. In order to allow a fair comparison with the baseline model, prediction ties—causes with the same predicted probability—were ordered similarly to the MV baseline model.
Configurations	μ	SET model parameter μ was set to zero. μ can be interpreted as the cause's general tendency to be active or inactive. With $\mu = 0$, none of the causes have a preference to be active or inactive.
	β	SET model parameter β. β can be interpreted as the general strength of all relationships between causes and errors. Different values were evaluated, using a search grid from 0 to 1.5, by $.1$. The β parameter affects calibration performance, but not ranking performance.
	a_{ij}	The individual edge weights in the SET model. In all configurations $a_{ij} = 0$ for causes that cannot produce a particular error. For causes that can produce a particular error, four different weight configurations were evaluated: unweighted ($a_{ij} = 1$ for all causes that can produce a particular error), weighted by the error distributions, weighted by the cause distributions, or weighted by the interaction between the latter two.
	Window Size	The number of consecutive error responses that are provided to the SET model. Five different window sizes were evaluated; windows with 1, 3, 7, 15, or 30 errors.
Metrics	MAP	The Mean Average Precision (MAP) takes the Average Precision (AP) for the top 1 to k predictions and provides the mean AP across students. This ensures that not only the proportion but also the order of relevant predictions are taken into account. The MAP is a metric for ranking performance. The R package *Metrics* was used to compute the MAP.
	Brier Score	Brier scores give the mean squared difference between the estimated cause probabilities and the potential causes of the observed error. The Brier score is a metric for calibration performance. The R package *DescTools* was used to compute Brier scores.

model. Finally, model performance was evaluated on the basis of the match between potential and expected causes.

The various configurations of the SET model, summarized in Table 4, were explored on a random selection of roughly 80% of the students in the data. The 20% holdout data was then

used to compare the best-performing SET model to the MV baseline model. The samples were stratified by the number of errors per student, to account for the severely skewed frequency distribution of error responses (see Figure 4). As the MV baseline model returns a ranking, but no probability estimates, we used the Mean Average Precision ranking metric to compare both models (see the Metrics section below).

2.4.3. Metrics

The SET and MV baseline models were compared with respect to their Mean Average Precision (MAP, Table 4), a metric particularly suited to recommendation systems such as the SET model. In this context, precision is intuitively defined as the ratio of the estimated causes *that are relevant to a user's potential causes* to all estimated causes (for a given observed error). Then, the Average Precision (AP) takes the average of the precision for the top one estimated cause up to the precision for the top n estimated causes. Finally, the MAP takes the mean of the AP for different windows and users. That is, the MAP maps the ranking of the predicted cause vector (ordered by estimated probability in the case of SET, or overall edge proportion in the case of MV) to the observed vector of potential causes, and quantifies this aspect of the quality of the model in a single value. The metric is defined as $\frac{\sum_{q=1}^{Q} AP(q)}{Q}$, where Q is the number of queries (a query is the set of predictions for a particular error window of a particular individual) and AP is the average precision. AP is defined as $\frac{\sum_{r=1}^{n} P(r) \times rel(r)}{\text{number of relevant documents}}$, where r is the rank of a prediction in the evaluated query, $P(r)$ is the precision at cut-off r, and $rel(r)$ is 0 if the prediction at rank r is not a possible cause of the evaluated error, and 1 otherwise. As generally one wants to consider only a small number of recommendations, we compute the MAP for different k, where k is the number of top predictions.

The MAP evaluates ranking accuracy and disregards the actual probability estimates, allowing a comparison of the SET model with the MV baseline model. However, the probability estimates do provide information on prediction reliability—or calibration—that is lost in the MAP analyses. Therefore, we computed Brier scores for the different evaluated configurations of the SET model in Table 4. Brier scores (Brier, 1950) are defined as $\frac{\sum_{r=1}^{N} (p_r - o_r)^2}{N}$ and give the mean squared difference between the estimated cause probabilities p and the potential causes o of the evaluated errors (0 if not a potential cause of the evaluated error, 1 otherwise), across the various elements r in a query.

2.4.4. Edge weights

For the SET model, we compared different methods of weighing the edges a_{ij} in the bigraph. In all configurations, $a_{ij} = 0$ for causes that cannot produce a particular error. For causes that can produce a particular error, the edges were either unweighted ($a_{ij} = 1$ for all causes that can produce a particular error), weighted by the distribution of the causes, or weighted by the distribution of the errors. First, the *Ockham's razor* hypothesis favors the simplest explanation of observed error responses. It states that the larger the number of errors a cause can explain the simpler the theory. The SET model with *unweighted* edges satisfies this hypothesis by means of the model's appeal to majority.

Second, the *law of diminishing returns* hypothesis retains the appeal to majority but diminishes its effect. The law of diminishing returns originates in the field of economics—but is also used in the fields of sports and learning—and signifies the diminishing returns of each additional

unit of for instance production (or practice in the case of sports and learning). In learning, the power law of practice is an explication of the law of diminishing returns for reaction times in practice trials. It states that with each additional trial, the return—in terms of the obtained reaction time—is diminished; a process that can be described by a power function. In the SET model, we used a power function to diminish the effect of the appeal to majority. We multiplied the edge weights of a cause by E^{-b}, where E is the number of errors in the bigraph that the cause can explain and $b = .5$. With $b = 0$ the edge weights would be unaltered, whereas with $b = 1$ the edge weights would effectively be divided by the number of error responses for each cause, such that the information is lost (i.e., becoming *unweighted*). This weight transformation ('*cause*' weights) ensured that for each additional error that is observed for a particular cause, the weights are diminished. The *cause weights* may affect calibration performance, but not ranking performance.

Third, the *specificity* hypothesis favors causes that uniquely explain particular error responses. It divides the edge weights of an error by the number of causes that can explain the error. It originates from the idea that an error that can only be explained by a single cause, provides more evidence for that cause than an error that can also be explained by other causes. This weight transformation ('*error*' weights) ensured that for each additional cause that can explain an observed error, the weights are diminished. The *error weights* may affect both calibration performance and ranking performance.

Finally, the edges were weighted by a simple interaction between the latter two weights; the *cause weights* times the *error weights*. These *cause*error weights* may affect both calibration performance and ranking performance.

3. Results

In this section, we first evaluate different configurations of the SET model on the training data, with respect to both ranking performance and calibration performance. Second, we compare the best performing SET model to the MV baseline model on the holdout data, with respect to their ranking performance. Finally, we confirm the calibration performance of the best performing SET model on the holdout data.

3.1. Training Data: SET Model Configurations

Figure 9 gives the Mean Average Precision at $k = 1, \ldots, K$ (MAP@k) for the various configurations of the SET model on the *training data*. Importantly, window size and edge weights (except for the *cause weights*) have an impact on the ranking of the predicted causes. Contrary to the window size and edge weights, k is not an optimization criterion, but rather reflects one's desired number of cause recommendations. From the figure, a clear pattern emerges where the MAP is higher for larger window sizes. A window size of 30 errors provides the highest MAP, regardless of the edge weight transformation and across all k. Then, with a window size of 30 errors, the *unweighted* model and model with *cause weights* perform best across all k.

Figure 10 gives the average Brier scores for the various configurations of the SET model on the *training data*. Brier scores allow for the evaluation of the SET model's probability estimates. Whereas the MAP solely considers ranking accuracy, Brier scores map the predicted probabilities to the potential causes of an observed error. Thus, whereas the *cause weights* and β parameter do not affect ranking performance, the various edge weight configurations, the β

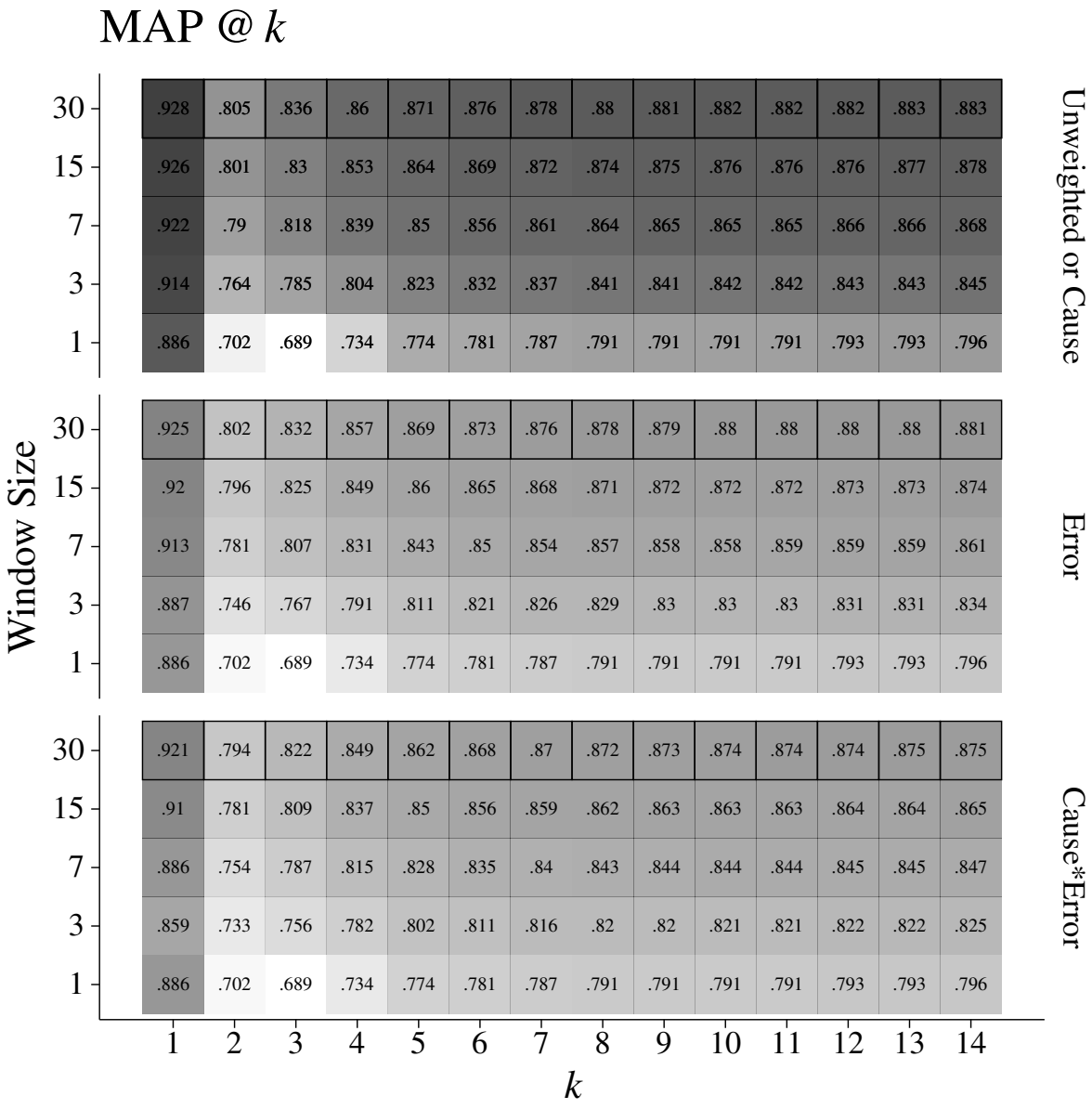

Figure 9: Heatmaps of the Mean Average Precision at k for the SET model, for different values of k (x-axis), different window sizes (y-axis), and weights (panels). Darker tiles represent higher values and thus improved ranking performance. Best performing configurations are highlighted for each k and edge weight transformation. Configurations were evaluated on the training data. Each data point is based on 278 students, where the average number of predictions per student varies between 51.6 and 80.6.

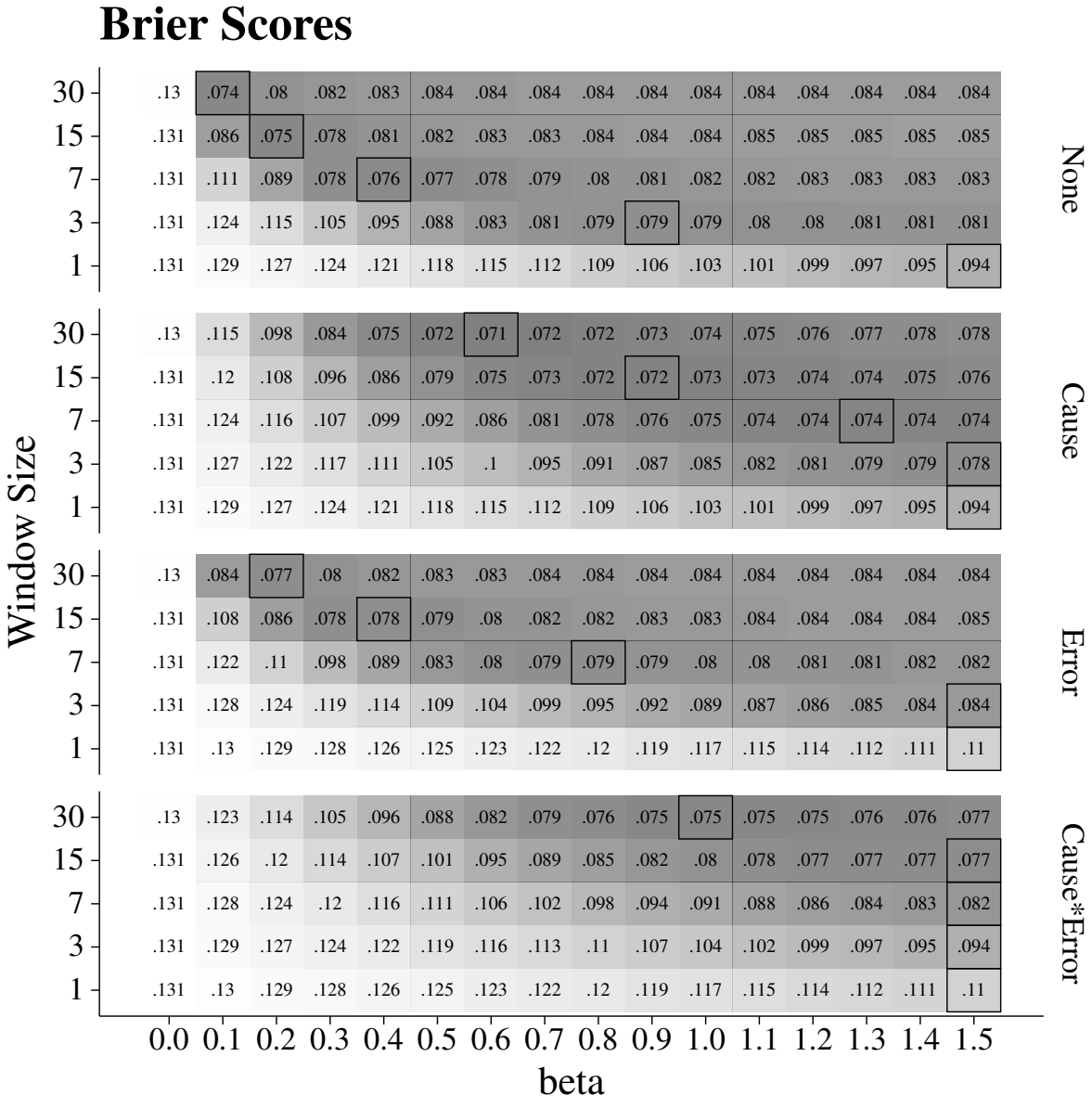

Figure 10: Heatmaps of the average Brier scores across predictions and students for the SET model, for different values of β (x-axis), different window sizes (y-axis), and weights (panels). Darker tiles represent lower values and thus improved calibration performance. Best performing configurations are highlighted for each window size and edge weight transformation. Configurations were evaluated on the training data. Each data point is based on 278 students, where the average number of predictions per student varies between 51.6 and 80.6.

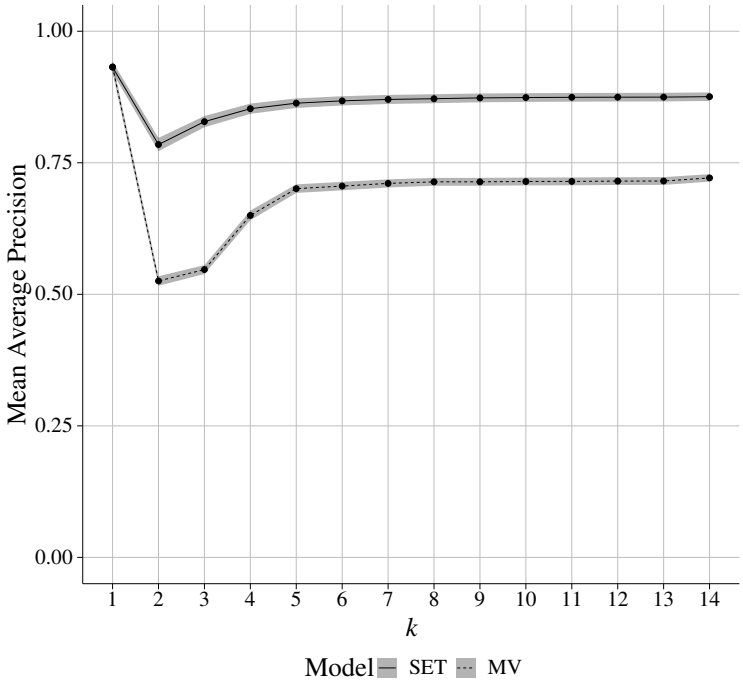

Figure 11: Mean average precision (y-axis) at k (x-axis) for the SET and MV baseline models (lines). Ribbons represent 95% confidence intervals. Models were evaluated on the holdout data. Each data point is based on 68 students, where the average number of predictions per student is 31.6.

parameter, and the window size may all affect calibration performance. From the figure, a clear interaction between window size and β emerges: smaller window sizes require larger β values. Again, a window size of 30 errors provides the lowest Brier score, regardless of the edge weight transformation. The overall lowest Brier score is obtained in the model with window size 30, *cause weights*, and $\beta = .6$.

3.2. Holdout Data: SET and MV Baseline Model Comparison

Figure 11 gives the MAP@k for the SET model and MV baseline model on the *holdout data*. For the SET model, the configuration that performed best in the training data was chosen; the model with window size 30, $\beta = .6$, and *cause weights*. To reiterate, β has no effect on the MAP, and *cause weights* do not alter the ranking with respect to the unweighted configuration. The figure shows that for $k = 1$ the MV baseline model and SET model perform equally well, and both reach a Mean Average Precision of over .9. The MV baseline model owes its excellent performance at $k = 1$ to the dominance of the *operator relevant* error category, evidenced in Figure 5. However, for $k > 1$, it is seen that the SET model outperforms the MV baseline model. Thus, whereas the models perform equally well if only the top recommendation is considered, the models diverge when the top $k > 1$ recommendations are considered in favor of the SET model.

An inspection of the cause profiles allows for a better understanding of these performance differences. Figure 12 compares the SET and MV baseline model with respect to the mean pre-

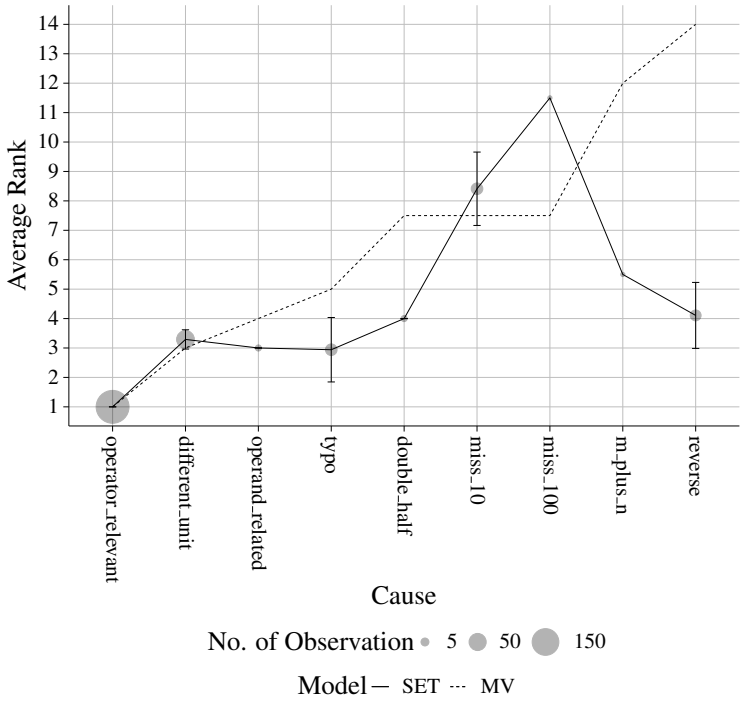

Figure 12: Average rank (y-axis) for the predictions of each potential cause (x-axis) observed in the data, for the SET and MV baseline models (lines). Error bars represent 95% confidence intervals. Point size represents number of observations. Causes are ordered by the rank order of the MV baseline model. Prediction ties were averaged. Lower ranks signal improved ranking performance. Models were evaluated on the holdout data.

dicted ranks of the causes. This allows one to inspect how well each of the individual potential causes is diagnosed by the models. Thus, whereas Figure 11 provides the overall performance, Figure 12 emphasizes the idiosyncrasies with respect to the predictions of the various causes. For each prediction, we determined the predicted ranks of the causes that were linked to the observed error response. Then, for every cause, we determined the average rank across all predictions. Prediction ties were averaged. The rank order of the MV baseline model is fixed by necessity and given in Table 2. Then, the figure suggests that the SET model provides substantially lower average ranks for 5 out of 9 causes and substantially higher average ranks for 2 out of 9 causes. However, the *operand related*, *double half*, *miss 100*, and *m plus n* causes have insufficient observations to permit a comparison with the baseline. Furthermore, the 95% confidence intervals of the *operator relevant*, *different unit*, and *miss 10* causes overlap with the baseline. The two remaining causes, the *typo* and *reverse*, do seem to perform substantially better than the baseline. The SET model thus seems to succeed in adapting its predicted ranking to students' error responses for the *typo* and *reverse* causes but fails to do so for others.

Finally, we determined the calibration performance of the SET model on the holdout data. The SET model that performed best in the training data had window size 30, *cause weights*, and $\beta = .6$. When evaluated on the holdout data (containing 68 students with on average 31.6 predictions per student), this configuration obtained a Brier score of .071; an exact replication of the evaluation on the training data.

4. DISCUSSION

To benefit from the information that is captured in students' errors, the causes of those responses must be determined. In domains where unique items are susceptible to most—or even all—known causes, and where most unique error responses can be explained by multiple causes, this challenge is not trivial. Nevertheless, in this paper, we showed that the Systematic Error Tracing model successfully extracts some of the signal that is captured in the students' error responses. Importantly, it outperforms a baseline model even though both capitalize on the same theoretical structure: the known relations between causes and errors. Ultimately, the SET model returns predicted probabilities for all considered causes, which may function as intervention recommendations for a tutor. In the following, we first discuss the evaluation of the SET model and second discuss its practical implications.

4.1. EVALUATION

Indeed, the SET model's recommendations provide a more accurate reflection of the successive errors of students than the baseline model's recommendations, with the exception of the precision of their top recommendation, for which both models perform equally well. The SET model's superior performance is thus due to the more precise alternative recommendations. Importantly, this overall improvement is reflected in the precision of the individual potential causes. The SET model ranks the predictions of most potential causes, when observed, equal to or higher than the baseline model would. It means that the tailoring of recommendations to individual students has the desired effect. This is particularly well reflected in the prediction precision of the *reverse* cause; a cause that only explains a small proportion of all possible errors, but is nevertheless frequently observed in Dutch primary education, as discussed in the introduction. Another potential cause for which the SET model outperforms the baseline model is the *typo*. Finally, the *miss 10* cause is poorly ranked by both models. The exact origin of the poorly performing *miss 10* cause, and the discrepancies between the (average) ranks for the *typo* and *reverse* causes in the SET and baseline model, are not easily inferred and must be elucidated in follow-up research. The most straightforward hypothesis would be that the latter two are more systematically produced by individual students than they are potential causes of the observed errors across all students. Vice versa, students may not systematically produce the former.

A devil's advocate may argue that the trivial baseline model is favored, as the top recommendations are paramount, and the alternative recommendations are not necessary for a recommendation system. This statement is problematic for several reasons. First, the baseline model is not as simple as it may seem, as it exploits a theoretical mapping of errors to error categories. Nevertheless, computationally it is free, as opposed to the SET model. Second, the exceptional performance of the baseline model at $k = 1$ is caused by the dominance of the *operator relevant* error category. However, this category inherits its dominance from its definition, not necessarily from its empirical prolificness. Third, the SET model manages to match the baseline's performance at $k = 1$, despite its improved prediction of less dominant error categories. Fourth, very much related and highly relevant, it enables individualization of recommendations, also at $k = 1$, contrary to the baseline model. Fifth, the individualization of the alternative recommendations is highly informative, as these may as well contain the actual cause, can be used to track the emergence or retreat of particular causes, and may be exploited for follow-up interventions, for instance, if an intervention on the top recommendation is unsuccessful.

In absolute terms, the SET model ranks its recommendations quite well, as evidenced by the

Mean Average Precision. Although it drops just below .8 when only the top two recommendations are considered, it remains well above it for all other numbers of considered recommendations. A Mean Average Precision of .8, when only the top two recommendations are considered, is for instance achieved when out of five samples of each two recommendations, four samples exclusively contain recommendations that match with the potential causes of an observed error, whereas one sample exclusively contains recommendations that do not match with the potential causes of an observed error. Moreover, the estimated average ranks for the individual potential causes seem to fluctuate between one and four. However, it must be noted that the *miss 10* cause is an exception to this rule, on average ranked at eight (similar to the MV baseline model), and the average ranks could not be estimated for all causes. Finally, the calibration of probability estimates for the recommendations on the holdout data matches that of the training data, with a Brier score of .071.

The performance patterns found for the practice data seem too robust to neglect, as those may signal implications that are generalizable beyond the application of single-digit multiplication. For one, taking into account more error responses to trace the potential causes of systematic errors generally tends to result in more precise rankings. Evidently, taking more error responses into account increases the power to determine the actual causes, and apparently, the causes generally do not change, even in the largest evaluated window size. However, this finding demands attenuation. The effect of decreasing the size of the error window seems to have an overall limited effect and might actually be beneficial in some circumstances. For instance, in less error-prone domains, the observed errors may span longer periods of time, such that expected changes in causes due to learning force one to take into account fewer errors. Following the same reasoning, the optimal window size could be student dependent. Students that practice irregularly may require smaller windows in order not to capture changes in errors due to learning. As such, larger window sizes carry the trade-off that time-intensive data must be available. Finally, smaller window sizes take us to a second pattern, observed in the calibration performance of the SET model. With fewer observed errors, higher values of β provide better performance, and as such may compensate for the smaller window size. Naturally, as the value of β has no effect on ranking, this only makes sense when the probability estimates are of interest for one's specific use case.

4.2. Implications

Altogether, these findings indicate that the model is very well suited for the identification of causes of systematic errors, and it has several important benefits. One benefit is that it is intuitive. The primary source of the model is the relative number of observed errors each cause can explain. Although from a prediction perspective one might not be concerned about the model being intuitive or not, in an educational context it is a clear advantage. Both students and teachers generally value an understanding of the origin of inferences like these, rather than having to deal with black box analytics. Also, the model can be relatively easily be implemented in online learning environments. Given that a theory-based bigraph that links causes to errors is constructed, the actual calculations are lightweight and may depend on a limited number of errors. And finally, in addition to being intuitive and lightweight, the model carries substantial weight as it is embedded in the well-understood Ising model (Kruis and Maris, 2016, for eg.).

Clearly, as outlined in the introduction, important challenges exist in identifying students' causes of error responses. The SET model requires a mapping model that links causes to errors; a

labor-intensive task to create, and we are not aware of areas in which this process was automated. On the other hand, many causes in many areas have been identified in the literature, and this subject-independent method allows one to collect these causes and—given the availability of appropriate data—calculate predictions for each of the causes. Also, future research could focus on learning the mappings from data, similar to Q-matrix learning (Liu et al., 2012).

Learning creates a challenge too. An individual's causes may naturally change over time, and indeed the whole purpose of diagnosis is to eliminate the causes of observed error responses. In other words, learning defeats stable causes, and the suitability of the chosen window of observed errors crucially depends on this stability. We solved this issue by continuously tracking causes over time, using a reasonable window size, and showed that the loss in prediction precision can partly be mitigated by the β parameter. However, to better understand the influence of the chosen window, a deeper understanding of developmental trends in causes and the typical severity of different causes across students would be valuable.

The fact that the model returns multiple predictions, ranked by their estimated probabilities, provides a pragmatic solution to learning on the one hand and prediction uncertainty on the other. In learning, the predicted probabilities of some causes may gradually increase over time, whereas those of others may gradually decrease. Also, in a recommendation system one may intervene on the top predicted cause (e.g., by providing additional instruction on a particular misconception), evaluate changes in the estimated cause probabilities, and if no meaningful changes occur, intervene on the second predicted cause.

Being aware of these benefits and challenges, a thorough understanding of the model and its implications demands a discussion of three important issues. To begin with, we analyzed the model with error responses from an adaptive learning environment. The primary reason for using this data was the large number of both students and responses, and we see no reason to believe that the model would perform differently with data from a non-adaptive learning environment. In addition, although adaptive data was used, the chosen domain contains a very homogeneous set of items, and the bigraph that captures the actual relations between causes and errors was specifically designed for the studied items. Within such a confined domain, one might safely assume that a student has a very limited amount of causes of systematic errors. However, when analyzing responses from a variety of domains the method may be less appropriate.

Second, the proposed model cannot take correct responses into account. One may view this as a shortcoming, arguing that correct responses can carry counter-evidence for certain causes. This is however not as straightforward as it may seem. A simple intuition is that a correct response invalidates any cause in the domain of interest. However, in the case of single-digit multiplication, this could mean that a single correct response would invalidate all causes. Obviously, this is not realistic and forces us to acknowledge that students may have localized causes, where some items are susceptible to their cause, whereas others are not (i.e., students may use different strategies for different groups of items, such as correct memory recovery for some, and an erroneous procedure for others). Determining these clusters for individual students is an interesting avenue for future research. Moreover, de Mooij et al. (2021) show that mouse trajectories leading up to correct responses may signal alternative response strategies that of course can be used in the SET model.

Very much related to this issue is the model's lack of memory. In its current form, the model does not take into account previous responses to the problems in an evaluated window. Did a student respond correctly to the problem before? How often did the student respond to the problem? Was there much variation in the responses to the problem? As such, previous

responses may provide valuable information about the stability of the potential causes. Future extension of the model may take such responses into account, although evidently it comes with a trade-off in degrees of freedom, as how much history should one for instance take into account?

Lastly, one might argue that the payoff of error analyses is limited. If a student makes errors, a teacher could simply provide additional instruction about the correct procedure, without the need to understand the specific cause. Yet interestingly, Muller et al. (2007a) and Muller et al. (2007b) argue that, in the domain of science learning, this method can have an undesirable effect. They first show that correct explanations may sometimes actually *reinforce* students' causes, and then show that discussing the cause as part of the instruction can make students aware of it. Although it is unclear to which domains these findings generalize, it serves as an important warning to not just blindly assume the benefit of instructions on solely the correct procedure. On top of that, identifying the exact cause can help select specific problems that target the cause, and provide additional tailored practice.

Following up on the findings of Muller et al. (2007a) and Muller et al. (2007b), we suggest viewing the diagnosis and treatment of errors as an actual instructional design principle. Contrary to errorless learning—the idea that learning does not benefit from errors—it should be acknowledged that systematic errors are inevitable, and that targeted diagnosis and treatment of these errors can really benefit the student. In such a diagnose-and-treat model of learning, learning can be described in terms of the elimination of systematic errors. Also, because of this focus on the causes of error responses, the instruction and practice that is subsequently provided will target what the student does not grasp, rather than what she or he already does.

4.3. Conclusions

The completed analyses are key in understanding the predictive performance of the model. Next, learning interventions can be executed on the basis of their recommendations. Interventions are not only an ultimate goal of error analyses—tailoring instruction or practice to the systematic error responses of a specific student—but also are a great tool in further determining its precision. Given that an intervention is effective for a given student with a given cause, its success reflects the precision of the model. The proof is in the pudding.

5. Acknowledgments

Alexander O. Savi, Department of Psychology, University of Amsterdam; Benjamin E. Deonovic, ACT, Inc., Iowa; Maria Bolsinova, ACT, Inc., Iowa; Han L. J. van der Maas, Department of Psychology, University of Amsterdam; Gunter K. J. Maris, ACT, Inc., Iowa, and Department of Psychology, University of Amsterdam.

B.E.D. is now at Corteva. M.B. is now at the Department of Methodology and Statistics, Tilburg University. G.K.J.M. is now at Tata Consultancy Services.

A.O.S., H.L.J.v.d.M, and G.K.J.M. were supported by Netherlands Organisation for Scientific Research (NWO) Creative Industries Grant 314-99-107. H.L.J.v.d.M. is full professor of Psychological Methods at the University of Amsterdam and founder of Oefenweb, the company that operated Math Garden.

Correspondence concerning this article should be addressed to Alexander O. Savi, Psychological Methods, Department of Psychology, University of Amsterdam, Nieuwe Achtergracht 129-B, 1018 WS Amsterdam, the Netherlands. Email: o.a.savi@gmail.com

REFERENCES

BEN-ZEEV, T. 1995. The nature and origin of rational errors in arithmetic thinking: Induction from examples and prior knowledge. *Cognitive Science 19,* 3, 341–376.

BEN-ZEEV, T. 1998. Rational errors and the mathematical mind. *Review of General Psychology 2,* 4, 366–383.

BRADSHAW, L. AND TEMPLIN, J. L. 2013. Combining item response theory and diagnostic classification models: A psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika 79,* 3, 403–425.

BRAITHWAITE, D. W., PYKE, A. A., AND SIEGLER, R. S. 2017. A computational model of fraction arithmetic. *Psychological Review 124,* 5, 603–625.

BRIER, G. W. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review 78,* 1, 1–3.

BRINKHUIS, M., SAVI, A., HOFMAN, A. D., COOMANS, F., VAN DER MAAS, H. L. J., AND MARIS, G. 2018. Learning as it happens: A decade of analyzing and shaping a large-scale online learning system. *Journal of Learning Analytics 5,* 2, 29–46.

BROWN, J. S. AND BURTON, R. R. 1978. Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science 2,* 2, 155–192.

BUWALDA, T., BORST, J., VAN DER MAAS, H. L. J., AND TAATGEN, N. 2016. Explaining mistakes in single digit multiplication: A cognitive model. In *Proceedings of the 14th International Conference on Cognitive Modeling*, D. Reitter and F. E. Ritter, Eds. 131–136.

CHEN, J. AND DE LA TORRE, J. 2018. Introducing the general polytomous diagnosis modeling framework. *Frontiers in Psychology 9*, 1474.

CONATI, C., GERTNER, A., AND VANLEHN, K. 2002. Using Bayesian networks to manage uncertainty in student modeling. *User Modeling and User-Adapted Interaction 12,* 4, 371–417.

CONWAY, J. R., LEX, A., AND GEHLENBORG, N. 2017. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics 33,* 18, 2938–2940.

CORBETT, A. T. AND ANDERSON, J. R. 1995. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modelling and User-Adapted Interaction 4,* 4, 253–278.

COWELL, R. G., DAWID, P., LAURITZEN, S. L., AND SPIEGELHALTER, D. J. 1999. *Probabilistic Networks and Expert Systems*. Information Science and Statistics. Springer-Verlag New York.

COX, D. R. AND WERMUTH, N. 1994. A note on the quadratic exponential binary distribution. *Biometrika 81,* 2, 403–408.

DE MOOIJ, S. M. M., RAIJMAKERS, M. E. J., DUMONTHEIL, I., KIRKHAM, N. Z., AND VAN DER MAAS, H. L. J. 2021. Error detection through mouse movement in an online adaptive learning environment. *Journal of Computer Assisted Learning 37,* 1, 242–252.

DESMARAIS, M. C. AND BAKER, R. S. J. D. 2011. A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction 22,* 1-2, 9–38.

GU, Y. AND XU, G. 2018. The sufficient and necessary condition for the identifiability and estimability of the DINA model. *Psychometrika 84,* 2, 468–483.

HAERTEL, E. H. 1989. Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement 26,* 4, 301–321.

ISING, E. 1925. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik 31,* 1, 253–258.

KHAJAH, M., LINDSEY, R. V., AND MOZER, M. C. 2016. How deep is knowledge tracing? In *Proceedings of the 9th International Conference on Educational Data Mining*, T. Barnes, M. Chi, and M. Feng, Eds. 94–101.

KLINKENBERG, S., STRAATEMEIER, M., AND VAN DER MAAS, H. L. J. 2011. Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education 57,* 2, 1813–1824.

KRUIS, J. AND MARIS, G. 2016. Three representations of the Ising model. *Scientific Reports 6,* 1, 1–11.

KUO, B.-C., CHEN, C.-H., AND DE LA TORRE, J. 2017. A cognitive diagnosis model for identifying coexisting skills and misconceptions. *Applied Psychological Measurement 42,* 3, 179–191.

KUO, B.-C., CHEN, C.-H., YANG, C.-W., AND MOK, M. M. C. 2016. Cognitive diagnostic models for tests with multiple-choice and constructed-response items. *Educational Psychology 36,* 6, 1115–1133.

LIU, J., XU, G., AND YING, Z. 2012. Data-driven learning of q-matrix. *36,* 7, 548–564.

LUCE, R. D. 2005. *Individual choice behavior: A theoretical analysis*. Dover Publications.

MARIS, G. AND VAN DER MAAS, H. L. J. 2012. Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika 77,* 4, 615–633.

MCCLOSKEY, M., HARLEY, W., AND SOKOL, S. M. 1991. Models of arithmetic fact retrieval: An evaluation in light of findings from normal and brain-damaged subjects. *Journal of Experimental Psychology: Learning, Memory, and Cognition 17,* 3, 377–397.

MITROVIC, A. 2011. Fifteen years of constraint-based tutors: what we have achieved and where we are going. *User Modeling and User-Adapted Interaction 22,* 1-2, 39–72.

MULLER, D. A., BEWES, J., SHARMA, M. D., AND REIMANN, P. 2007a. Saying the wrong thing: improving learning with multimedia by including misconceptions. *Journal of Computer Assisted Learning 24,* 2, 144–155.

MULLER, D. A., SHARMA, M. D., EKLUND, J., AND REIMANN, P. 2007b. Conceptual change through vicarious learning in an authentic physics setting. *Instructional Science 35,* 6, 519–533.

NORMAN, D. A. 1981. Categorization of action slips. *Psychological Review 88,* 1, 1–15.

PEARL, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

PELÁNEK, R. 2017. Bayesian knowledge tracing, logistic models, and beyond: An overview of learner modeling techniques. *User Modeling and User-Adapted Interaction 27,* 3-5, 313–350.

PIECH, C., BASSEN, J., HUANG, J., GANGULI, S., SAHAMI, M., GUIBAS, L. J., AND SOHL-DICKSTEIN, J. 2015. Deep knowledge tracing. In *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 505–513.

REBER, R., BRUN, M., AND MITTERNDORFER, K. 2008. The use of heuristics in intuitive mathematical judgment. *Psychonomic Bulletin & Review 15,* 6, 1174–1178.

SHAKI, S. AND FISCHER, M. H. 2017. Competing biases in mental arithmetic: When division is more and multiplication is less. *Frontiers in Human Neuroscience 11*, 37.

STRAATEMEIER, M. 2014. Math garden: A new educational and scientific instrument. Ph.D. thesis.

TARAGHI, B., FREY, M., SARANTI, A., EBNER, M., MÜLLER, V., AND GROSSMANN, A. 2015. Determining the causing factors of errors for multiplication problems. In *Communications in Computer and Information Science*. Springer International Publishing, 27–38.

TARAGHI, B., SARANTI, A., LEGENSTEIN, R., AND EBNER, M. 2016. Bayesian modeling of student misconceptions in the one-digit multiplication with probabilistic programming. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*. ACM Press, 449–453.

TATSUOKA, K. K. 1983. Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement 20,* 4, 345–354.

TEMPLIN, J. L. AND HENSON, R. A. 2006. Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods 11,* 3, 287–305.

VAN DER VEN, S. H. G., KLAIBER, J. D., AND VAN DER MAAS, H. L. J. 2017. Four and twenty blackbirds: How transcoding ability mediates the relationship between visuospatial working memory and math in a language with inversion. *Educational Psychology 37,* 4, 1–24.

VANLEHN, K. 1986. Arithmetic procedures are induced from examples. In *Conceptual and procedural knowledge: The case of mathematics*, J. Hiebert, Ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 133–179.

VOMLEL, J. 2004. Bayesian networks in educational testing. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 12,* supp01, 83–100.

XU, Y. AND MOSTOW, J. 2011. Using logistic regression to trace multiple sub-skills in a dynamic Bayes net. In *Proceedings of the 4th International Conference on Educational Data Mining*, M. Pechenizkiy, T. Calders, C. Conati, S. Ventura, C. Romero, and J. Stamper, Eds. 241–246.

APPENDIX

A. SLIPS

Using the model outlined in the Methods section, and the method outlined in Equation 6 to quantify the likelihood of a particular cause, it is irrelevant whether the model includes a 'slip' cause that is connected to every error. We will show that the probability that cause i and a slip is present, given that cause i is the only additional cause, is the same as the probability that cause i is present, given only one cause is present, in the model without slips. Let μ_s be the external field parameter associated with the slip and $a_{sj} = 1$ for all $j = 1, \ldots, m$ to indicate the slip is connected to all errors.

$$p_{is} = p(c_i = 1, c_s = 1 | \textstyle\sum_{i=1}^{n} c_i = 1, c_s = 1, \mathbf{e} = \mathbf{1}, \boldsymbol{\mu}, \beta) \tag{10}$$

$$= \frac{\exp\left(\beta\mu_i + \beta\mu_s + \beta\sum_{j=1}^{m}\mu_{n+j} + 2\beta\sum_{j=1}^{m}a_{ij} + 2\beta\sum_{j=1}^{m}a_{sj}\right)}{\sum_{k=1}^{n}\exp\left(\beta\mu_k + \beta\mu_s + \beta\sum_{j=1}^{m}\mu_{n+j} + 2\beta\sum_{j=1}^{m}a_{kj} + 2\beta\sum_{j=1}^{m}a_{sj}\right)} \tag{11}$$

$$= \frac{\exp\left(\beta\mu_i + \beta\sum_{j=1}^{m}\mu_{n+j} + 2\beta\sum_{j=1}^{m}a_{ij}\right)\exp\left(\beta\mu_s + 2\beta\sum_{j=1}^{m}a_{sj}\right)}{\sum_{k=1}^{n}\exp\left(\beta\mu_k + \beta\sum_{j=1}^{m}\mu_{n+j} + 2\beta\sum_{j=1}^{m}a_{kj}\right)\exp\left(\beta\mu_s + 2\beta\sum_{j=1}^{m}a_{sj}\right)} \tag{12}$$

$$= \frac{\exp\left(\beta\mu_i + \beta\sum_{j=1}^{m}\mu_{n+j} + 2\beta\sum_{j=1}^{m}a_{ij}\right)}{\sum_{k=1}^{n}\exp\left(\beta\mu_k + \beta\mu_s + \beta\sum_{j=1}^{m}\mu_{n+j} + 2\beta\sum_{j=1}^{m}a_{kj}\right)} \tag{13}$$

$$= p(c_i = 1 | \textstyle\sum_{i=1}^{n} c_i = 1, \mathbf{e} = \mathbf{1}, \boldsymbol{\mu}, \beta) \tag{14}$$

$$= p_i \tag{15}$$

Crucially, it is assumed that at least one actual cause is present. This assumption is justified by the selection of the error responses. Our data selection criteria are aimed at filtering out the individuals that provide systematic error responses, by means of constraints on the minimum number of observed error responses within a particular time window. The inherent nature of slips renders it unlikely to observe many slips for a single individual in a constrained time window. In the case that our filter fails, we are confident that the obtained recommendations—although wrong—do not hurt.