Exceptional Subitizing Patterns: Exploring Mathematical Abilities of Finnish Primary School Children with Piecewise Linear Regression

Rianne M. Schouten(✉)[1], Wouter Duivesteijn[1], Pekka Räsänen[2], Jacob M. Paul[3], and Mykola Pechenizkiy[1]

[1] Eindhoven University of Technology, the Netherlands
{r.m.schouten,w.duivesteijn,m.pechenizkiy}@tue.nl
[2] Turku Research Institute for Learning Analytics, Faculty of Science, University of Turku, Finland, pekka.j.rasanen@utu.fi
[3] School of Psychological Sciences, University of Melbourne, Australia, jacob.paul@unimelb.edu.au

Abstract. Numerical processing competences such as the ability to enumerate small sets of dots and to compare the relative magnitudes between sets are diagnostic markers of young children's emerging math abilities. In the FUnctional Numerical Assessment (FUNA) study, these abilities are assessed using several computer-assisted tasks, among which is a Dot Enumeration (DE) task where children determine the number of dots in a visual array. It seems that there is a natural threshold around 3 or 4 dots: below this threshold, it is possible to determine the correct number at a glance, known as subitizing; above the threshold, children must count the dots in some way. In this paper, we develop a piecewise linear regression model class for Exceptional Model Mining with various quality measures discovering subgroups of children whose subitizing curves exhibit atypical patterns. The dataset does not follow the conventional data mining representation where each individual is described with a tuple of attribute values. Rather, for each task, students perform multiple items, one after the other, taken from a larger set of items, and not necessarily in the same order. Hence, we discuss a manner (tailored to the dataset at hand) to transform this item-performance data into the flat-table form that the typical data mining task expects. Domain experts confirm that our experiments evidently demonstrate how children's subitizing performance and counting skills are related to math abilities. Our findings provide opportunity for further development of assessment tools and intervention programs.

Keywords: Learning Analytics · Exceptional Model Mining · Piecewise Linear Regression

1 Introduction

Learning mathematics is hard. At the neuro-cognitive foundation of young children's math development are core numerical processing competences such as the

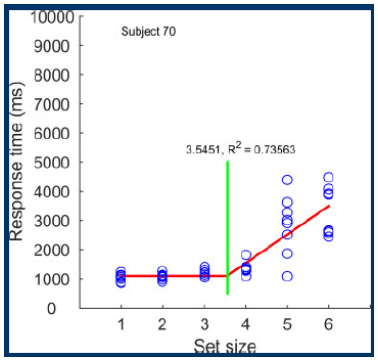

Fig. 1. Dot enumeration response time regressed on set size (number of dots) using segmented linear regression with one break point.

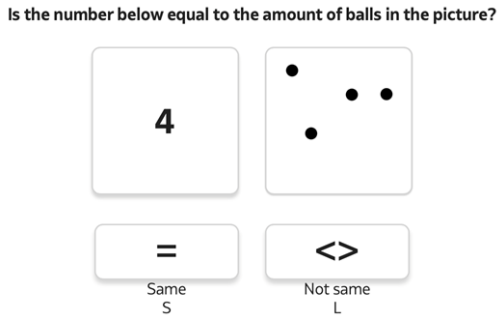

Fig. 2. An example of a dot enumeration item for set size 4; the correctly answered items are used in Figure 1.

ability to enumerate small sets of dots and to compare the relative magnitudes between sets [6]. These numerical competences are diagnostic markers of emerging math abilities from as early as preschool age [7] which make them targets for conceptually motivated intervention programs [1].

We investigate characteristics that define exceptional patterns of young children's enumeration ability. Generally, enumeration performance reflects two distinct processes: the *subitizing* system where small sets (1-4 dots) are recognized accurately and rapidly, and the *counting* system where larger sets are enumerated more slowly perhaps by counting or other enumeration strategies [23]. Figure 1 gives an example; the enumeration response time of small sets is relatively flat while the counting slope is steeper. The inflection point demarcates the subitizing range from the counting range.

Individual differences in subitizing range predict math ability [17]. An inability to subitize is associated with dyscalculia [12]. There is value in accurately and reliably estimating the parameters that define subitizing patterns (initial reaction time, range, slope). Common algorithms used for estimating the subitizing range can produce inconsistent results [15], especially among individuals with dot enumeration curves that deviate from the typical curve.

We develop a piecewise linear regression model class for Exceptional Model Mining (EMM) [4] to discover subgroups of children whose subitizing curves exhibit atypical patterns. EMM is a local pattern mining framework seeking coherent subgroups of a dataset that somehow behave exceptionally. We develop various quality measures based on the concept of log likelihood that allow us to discover atypical subitizing patterns such as deviating initial reaction times, subitizing ranges, counting slopes, or a combination of those.

We use data collected by the FUnctional Numerical Assessment (FUNA) study [22]. Numerical processing competences and math abilities are assessed using several computer-assisted *tasks*. Some of these tasks contain a fixed num-

ber or questions, or *items*; others are time-based and the number of answered items will vary per child. Items are taken from a larger set of items, and not necessarily answered in the same order. Consequently, the dataset does not follow the conventional data mining representation where each individual can be described with one tuple of attribute-values, and where a column contains the same semantic information for each individual[4]. Hence, pre-processing is required to allow existing algorithms to search through the space of candidate subgroups. We discuss a manner tailored to the item-performance data at hand.

The main contributions of this paper are: 1) an EMM model class and various quality measures for segmented linear regression; 2) a deeper understanding of how subitizing patterns relate to other numerical processing competences and emerging math abilities; 3) an effective pre-processing technique for handling repeatedly measured attributes in descriptive space.

2 The FUnctional Numerical Assessment study

The FUnctional Numerical Assessment (FUNA) project [22] is a large-scale research program in Finland to develop digital assessment tools for detecting dyscalculia and dyslexia. Currently, several studies are run to evaluate the validity and reliability evidence of the tasks [8]. The current version has been normed in Finnish and Finnish-Swedish languages for grade levels 3 to 9 (9 to 15 years old). In the FUNA-DB (Dyscalculia Battery) the children respond to six digital (CAI) *tasks* using a tablet or a computer: Number Comparison (NC), Dot Matching equivalence task (DM), Single Digit Addition (SA), Single Digit Subtraction (SS), Combination Addition (CA) and Number Series (NS). Every task consists of multiple questions, or *items*. The tasks SA, SS, CA, and NS measure arithmetic fluency, and the items considered easier are provided earlier than more difficult items, but the exact order is not the same between children (i.e. quasi-random). In the number processing tasks (NC, DM), a set of predefined items are presented in a fully random order. Figure 2 displays an example of a DM item. Children compare a symbolic number (1-9) to a non-symbolic representation of a number. The location of the dots is randomized as well. When the symbolic and non-symbolic representations are the same, and when the children answer correctly, the DM task can be considered a Dot Enumeration (DE) task: determining the number of dots in a visual array.

Table 1 displays a dataset slice. On the right side (to be used as *target* attributes in the EMM model class, see Section 5), we present information from the DE task. Attributes ℓ_1 and ℓ_2 represent the set size (1-9) and response time

[4] Children build up experience with the type of tasks at hand while the study unfolds. Suppose that two children perform Task T, but Child A is given this task earlier in the procedure than Child B. Then, Child B will have built up more experience than Child A with similar tasks, before executing Task T. A conventional data mining representation of this data would record performance of both children on Task T in the same column, but this belies the reality that these performances are not measured in an equal manner.

Table 1. Small slice of FUNA dataset. Some descriptors originate from the NC-task (a_2,a_3,a_4), others from the SA (a_5), SS, or CA task, or from the general background information (sex, a_1). In our EMM instance, target attributes originate from the DE task (ℓ_1, ℓ_2). All task-based attributes contain data from multiple items, resulting in tuples of values. The number of values per tuple may vary per child and per task.

	sex	NC			SA		DE	
i	a_1	a_2	a_3	a_4	a_5	...	ℓ_1	ℓ_2
1	f	(4,3,1,4,...)	(1,0,0,...)	(1200,1150,...)	(6,2,2,...)	...	(5,1,8,...)	(1330,14...)
2	f	(2,3,1,7,...)	(1,1,1,...)	(1240,1510,...)	(5,2,4,...)	...	(8,2,1,...)	(2630,21...)
3	m	(5,2,8,6,...)	(0,1,1,...)	(1490,1250,...)	(7,3,1,...)	...	(4,9,2,...)	(2130,19...)
4	f	(8,2,1,5,...)	(0,1,1,...)	(1180,1120,...)	(3,7,7,...)	...	(7,4,5,...)	(2610,16...)

in milliseconds respectively. These attributes are the dependent and independent variables in a segmented linear regression model class as visualized in Figure 1. We indicate the fact that we obtain data from multiple DE items per child, by using tuples (e.g., for the first item of child 1, the set size was 5 and response time was 1330 ms). For the SA, SS, and CA tasks, the number of items (tuple-length) differs per child; for the NC and DE tasks, the tuple-length is 52.

Apart from the set size and response time for each task, we may consider information such as whether the item is answered correctly, what is the correct answer, and what is the numerical distance between two numbers shown in a certain item. All this information is represented as separate attributes (e.g., attribute a_3 indicates where the items on the NC task have been answered correctly (1) or not (0)) and will be used to discover and *describe* exceptional subgroups of children. We also have some descriptive information, such as a child's sex (a_1), grade, and the language (Finnish or Swedish) in which they executed the tasks.

The data format as used by most traditional data mining algorithms is also known as a propositional table; these are single-table representations where each individual can be described with one term. In the attribute-value case, this term is a tuple of attribute values [13]. For instance, a student could be represented by a three-tuple specifying age, grade and language. Generally in EMM, we let the subgroup description be a conjunction of selection conditions over the descriptors, where condition sel_j is a restriction on the domain $\mathcal{A}_j$ of the respective attribute a_j. For instance, a description sex = girl $\wedge$ language = Finnish covers all girls who executed the FUNA tasks in Finnish.

However, for all attributes other than sex, grade and language, our dataset does not follow this conventional data mining representation; a descriptive attribute is not associated to one value, but rather to a tuple of values. In this case, it is unclear what it means to apply a selector sel_j directly; a selector $a_2 \leq 3$ would select items rather than individuals and a selector such as $a_{2t} \leq 3$ where t refers to the item indicator, would inflate the number of descriptors, which is detrimental to efficient traversal of the search space. In addition, such a selector has little conceptual meaning, again because the items are quasi-randomly ordered and item t is not the same across children. We will provide a more satisfactory alternative in Section 4.

3 Background

Exceptional Model Mining (EMM) [4] is a local pattern mining framework seeking coherent subgroups in the dataset that somehow behave exceptionally. The observed attribute-values are divided into descriptors $a_1, \ldots, a_k$ and targets $\ell_1, \ldots, \ell_m$. Dataset Ω is then a bag of n records $r \in \Omega$ of the form

$$r = (a_1, \ldots, a_k, \ell_1, \ldots, \ell_m). \tag{1}$$

Subgroups are defined using descriptions; a Boolean function $D : \mathcal{A} \to \{0, 1\}$. A description D covers a record r^i if and only if $D(a_1^i, ..., a_k^i) = 1$.

Definition 1 (Subgroup cf. [4]). *A subgroup corresponding to description D is the bag of records $G_D \subseteq \Omega$ that D covers:*

$$G_D = \{r^i \in \Omega \mid D(a_1^i, a_2^i, \ldots, a_k^i) = 1\}.$$

The complement contains all non-covered records: $G_D^C = \Omega \setminus G_D$ [4, p.53].

In EMM, the choice of description language $\mathcal{D}$ is free, though generally we let the description be a conjunction of selection conditions over the descriptors, where condition sel_j is a restriction on the domain $\mathcal{A}_j$ of the attribute a_j. For discrete variables the selector may be an attribute-value pair $(a_j = v)$; for continuous variables it could be a range of values $(w_1 \leq a_j \leq w_2)$ [4].

The task of EMM is to discover the descriptions whose subgroups display exceptional behaviour on the target variables. The precise instantiation of "behaviour" depends on the application. A quality measure quantifies the exceptionality of within-subgroup behaviour with some reference behaviour model.

Definition 2 (Quality Measure cf. [4]). *A quality measure (QM) is a function $\varphi : \mathcal{D} \to \mathbb{R}$ that assigns a numerical value to a description D.*

The challenge in EMM is to effectively search through the descriptive space to find the top-q best-scoring subgroups.

In traditional EMM, the combination of Equation (1) and a description language based on conjunctions of selection conditions implicitly assumes the data to be in a flat-table format where every record is an individual that is described by a tuple of attribute values, and placed on a new row in the single flat-table. In contrast, in this paper, an attribute a or ℓ may or may not be measured repeatedly per individual i. We focus our notation on the descriptive attributes, and write a_{jt}^i to denote the t^{th} measurement of the j^{th} descriptive attribute for the i^{th} individual. We use the term *record* to refer to an *individual*; compared to Equation (1), the form of the descriptive part of record $r \in \Omega$ changes to:

$$r = ((a_{11}, a_{12}, ..., a_{1t}, ..., a_{1t_1}), (a_{21}, ..., a_{2t_2}), ..., (a_{k1}, ..., a_{kt_k})), \tag{2}$$

where t_j^i refers to the number of repeated measures of attribute a_j for individual $i \in \{1, 2, ..., n\}$, which may vary across individuals and attributes; we let $t_j = \max_{i=1,2,\ldots,n} t_j^i$. Some descriptors may be measured only once per individual (such as sex in Table 1); then, $t_j^i = 1$ for all i.

3.1 Segmented linear regression

The goal of regression is to predict the value of an attribute y given a new value of $\mathbf{x}$, where $\mathbf{x}$ is a random draw from a vector of variables $\mathbf{X} = (X_1, \ldots, X_d)$. The simplest linear model for regression is one that involves a linear combination of the input variables and parameters $\mathbf{w}$: $f(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T\mathbf{x}$. We additionally aim to model the uncertainty, modeling a predictive distribution $p(y|\mathbf{x})$ by assuming that the deterministic function $f(\mathbf{x}, \mathbf{w})$ has additive Gaussian noise with zero mean and precision β (inverse variance). We then obtain the likelihood function:

$$p(y|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(y|f(\mathbf{x}, \mathbf{w}), \beta^{-1}) = \prod_{i=1}^{n} \mathcal{N}(y[i] \mid \mathbf{w}^T\mathbf{x}[i], \beta^{-1}), \tag{3}$$

Next, estimating $\mathbf{w}$ and β using Maximum Likelihood Estimation shows that the log likelihood of a regression model depends on the sum-of-squares error function (SSR) [2] (see [25, Section 1] for an elaboration):

$$\ln p(y|\mathbf{x}, \mathbf{w}, \beta) \approx SSR(y, f(\mathbf{x}, \mathbf{w})) = \sum_{i=1}^{n} \left(f(\mathbf{x}[i], \mathbf{w}) - y[i]\right)^2.$$

Segmented linear regression appears to require non-standard optimization techniques. However, one can parameterize the model such that it can be modeled using an iterative, linear approach [18]. We focus on modeling a segmented relationship with two line segments between response variable y and one explanatory variable x_h by fitting the terms:

$$y = g(x_h, \alpha, \beta, \psi) = \alpha x_h + \beta(x_h - \psi)_+ \tag{4}$$

where $(x_h - \psi)_+ = (x_h - \psi) \cdot I(x_h > \psi)$ where $I(\cdot)$ is the indicator function equal to 1 if the statement is true and 0 otherwise. Consequently, ψ is the x-axis break point, α is the slope of the line segment to the left of ψ, and β is the difference in slopes between the line segments to the left and right of ψ. Next, [18] iteratively fit linear models of the form $\alpha x_h + \beta U^{(s)} + \gamma V^{(s)}$ with $U^{(s)} = (x_h - \psi^{(s)})_+$ and $V^{(s)} = -I(x_h > \psi^{(s)})$. Every iteration, $\hat{\psi}^{(s+1)}$ is updated through $\left(\psi^{(s+1)} - \psi^{(s)}\right) = \hat{\gamma}/\hat{\beta}$ and when the algorithm stops and $\hat{\gamma} \approx 0$, the s^{th} approximation is the Maximum Likelihood Estimate: $\hat{\psi}^{(s)} \equiv \hat{\psi}$ [18].

3.2 Connections to existing SD/EMM approaches

Linear target models for EMM are not a new concept [20]. Existing model classes use QMs comparing a regression parameter between the subgroup and a reference model. Instead, we follow the approach of [27] and [24] who build QMs on the log likelihood. These QMs do not directly compare parameter estimates but rather evaluate the overall fit of a model estimated on the subgroup. In addition, in this paper, we utilize the special situation that when we assume Gaussian noise, maximizing the log likelihood is similar to minimizing the residuals sum-of-squares. This characteristic simplifies the notation and calculation of our QMs.

Our dataset has a nested structure: we aim to create subgroups at the level of the individual, while having access to repeated measures per individual in both target and descriptive space. We are not the first to consider time-varying target attributes. For instance, [24] analyzed glucose fluctuations and [3] discovered funding applications with deviating temporal sub processes. However, in descriptive space, these authors use attributes that are measured at the same level as the individual; their flattening approach can be categorized as a transformation to a wide flat-table data format. Alternatively, [16] transformed their data into a long, stacked flat-table format where each rows contains a transition rather than an entire sequence. Under the hood, some form of *propositionalization* [13] takes place in [3,16,24], transforming hierarchical data with one-to-many relations into a single-table representation where each individual can be described with one term. For [16], this involved a change of the notion of an individual.

Relational subgroup discovery (RSD) [28] uses a proportionalization-based approach that accepts feature language declarations similar to those used in Progol [19]. Our proposed method is best described as a simple *aggregation* approach to feature construction [11]. We do not apply automated feature construction methods; these typically assume that columns of the dataset have a coherent semantic meaning, which our data does not (cf. Footnote 1). We show that with domain-specific aggregation functions, subgroup interpretability blossoms.

4 Our proposed flattening approach

An *aggregated descriptor* is a descriptive attribute constructed out of one or more original descriptors, where the original descriptors are defined as in Section 3 and may or may not contain repeated measures per individual. The goal is to describe each individual with one tuple of attribute-values as in Equation (1), rather than a tuple of tuples as in Equation (2). This allows defining descriptions as conjunctions of selection conditions over the aggregated descriptors.

Denoting an original descriptor with a_j, we construct an aggregated descriptor $\tilde{a}_h$ by applying a function $\xi : \mathbb{R}^* \rightarrow \mathbb{R}^1$ such that per individual, the number of observed values on attribute $\tilde{a}_h$ is 1. A function ξ may be applied to one or more time-varying descriptors, possibly in combination with an invariant descriptor.

Definition 3 (Aggregated descriptor). *Given one or more descriptors $a_* \subseteq \{a_1, a_2, ..., a_k\}$, an aggregated descriptor $\tilde{a}_h$ is an attribute constructed by applying a function $\xi : \mathbb{R}^* \rightarrow \mathbb{R}^1$ such that per individual, the number of observed values on attribute $\tilde{a}_h$ is 1, i.e.: $\tilde{a}_h = \xi(a_*)$ with $a_* \subseteq \{a_1, a_2, ..., a_k\}$.*

Aggregated descriptors may arise from a function such as a summation or average, they may be non-linear (conditional) functions of one or more original descriptors, and/or they could be parameter estimates of a statistical model. Section 4.1 provides examples of all of these for the FUNA study.

The aggregated descriptors induce a tweak to the definition of a subgroup:

Table 2. An overview of the aggregation functions used in FUNA.

Tasks	Name	Explanation
NC,SA,SS,CA	MaxItem	Number of answered items
NC,SA,SS,CA	SumAnsC	Number of correctly answered items
NC,SA,SS,CA	PropAnsC	Proportion of correctly answered items
NC,SA,SS,CA	MeanTC	Mean response time of correctly answered items
NC,SA,SS,CA	MedTC	Median response time of correctly answered items
NC,SA,SS,CA	IES	Inverse Efficiency Score
NC	IcNumD	Intercept of the response time regressed on the distance between the two numbers of every item
NC	SlNumD	Slope of the response time regressed on the distance between the two numbers of every item
NC	IcNumR	Intercept of the response time regressed on the ratio between the distance and the largest of the two numbers of every item
NC	SlNumR	Slope of the response time regressed on the ratio between the distance and the largest of the two numbers of every item

Definition 4 (Subgroup). *A subgroup corresponding to description D is the bag of records $G_D \subseteq \Omega$ that D covers:*

$$G_D = \{r^i \in \Omega | D(\tilde{a}_1, \tilde{a}_2, ..., \tilde{a}_s) = 1\}. \tag{5}$$

The descriptive domain $\tilde{\mathcal{A}}$ is the collective domain of all aggregated descriptors $\tilde{a}_1, \ldots, \tilde{a}_s$ and the time-invariant descriptors $a_\dagger = \{a_j \in \{a_1, \ldots, a_k\} | t_j = 1\}$.

4.1 Domain-specific aggregations functions

Definition 3 allows for many variations. In the context of FUNA, a simple example is a function $\xi_{\max}$ that counts the number of answered items per task. For instance, $\tilde{a}_1^i = \xi_{\max}(a_{\text{NC}}^i) = t_{\text{NC}}^i$ is the number of NC items answered by individual i, where a_{NC} is the item-indicator of task NC. We may want to know how many items individual i answered correctly: $\tilde{a}_2^i = \xi_{\text{sum}}(a_3^i) = \sum_t a_{3t}^i$, where a_3 is a binary attribute as in Table 1. We could subsequently measure the proportion of correctly answered NC items: $\tilde{a}_3^i = \xi_{\max}(a_{\text{NC}}^i)/\xi_{\text{sum}}(a_3^i)$.

Other aggregation functions that are interesting from a domain perspective are the mean and median response time of the correctly answered items. We write $\tilde{a}_4^i = \xi_{\text{meanTC}} = (\xi_{\text{sum}}(a_3^i))^{-1} \cdot \sum_{t \in \{1,\ldots,t_4\} \text{ s.t. } a_{3t}^i = 1} a_{4t}^i$. For ξ_{medianTC} we would do something similar but take the median rather than the mean.

In the domain of educational learning, the Inverse Efficiency Score (IES) [7] is a measure that combines both the median response time and the accuracy (proportion of correctly answered items). The IES allows researchers to identify children with high response times, or a low proportion of correctly answered items, since the IES score is high in both cases. For an individual:

$$\tilde{a}_6^i = \xi_{\text{IES}}(a_{\text{NC}}^i, a_3^i, a_4^i) = \frac{\xi_{\text{MedianT}}(a_4^i)}{\xi_{\text{PropAnsC}}(a_{\text{NC}}^i, a_3^i)} = \frac{1/t_4^i \sum_t a_{4t}^i}{t_{\text{NC}}^i / \sum_t a_{3t}^i}. \tag{6}$$

For the Number Comparison (NC) task, it is interesting to analyze the numerical distance effect [9]. When tasked with saying which of two numbers is greater, this task is easier to perform when the numbers are far apart (*NumD*). If numbers have the same distance, the task is hypothesized [23] to be easier if the largest number is smaller. This is called the Number Ratio (*NumR*). We regress the response time of the NC items on the NumD (and once more for NumR), and evaluate the intercept (Ic) and Slope (Sl) of these models. Thus, we first create a time-variant descriptor $a_{\text{NumD}} = |a_{\text{NCL}} - a_{\text{NCR}}|$ (where a_{NCL} and a_{NCR} are the numbers shown on the left and right in each NC item) and then fit a linear regression model per individual i: $a_4^i = f(a_{\text{NumD}}^i, w_0^i, w_1^i)$. Parameter estimates w_0^i and w_1^i are the intercept and slope of the regression model, stored as aggregated descriptors $\tilde{a}_7^i = w_0^i$ and $\tilde{a}_8^i = w_1^i$. We take the same approach for NumR.

An overview of these aggregated descriptors is given in Table 2.

5 Our proposed target model

We seek subgroups of children with atypical dot enumeration curves. We use the segmented linear regression model as a target model (cf. Section 3.1) with response time ℓ_2 as output (y) and set size ℓ_1 as input (x_h) (cf. Table 1). We are interested in finding any kind of deviation from the typical DE curve; in a typical DE curve the subitizing slope is close to zero, the subitizing range is somewhere between 3 and 4, and the counting slope is relatively steep.

Following [27] and [24], we assume that the parameters of a linear model fitted on the subgroup will likely describe the subgroup better than the parameters estimated on the entire dataset. Then, in the presence of a subgroup, the log likelihood of dataset Ω will increase if the parameters of the subgroup are separately estimated. For any subgroup SG and its complement SG^C,

$$\ln p(SG|\theta^{SG}) + \ln p(SG^C|\theta^{\Omega}) > \ln p(SG|\theta^{\Omega}) + \ln p(SG^C|\theta^{\Omega}),$$

where $\ln p(SG|\theta^{SG})$ is the log likelihood of the subgroup for a segmented linear regression model estimated on the SG with $\theta^{SG} = (\alpha^{SG}, \beta^{SG}, \psi^{SG})$. We expect this term to be larger than the log likelihood of the subgroup for a segmented linear regression model estimated on the entire dataset Ω: $\ln p(SG|\theta^{SG}) > \ln p(SG|\theta^{\Omega})$. Next, we use the characteristic of linear regression that maximizing the log likelihood is similar as minimizing the sum-of-squares error function (SSR) (see Section 3.1, and [25, Section 1]) and aim to find subgroups where $\ln p(SG|\theta^{SG}) > \ln p(SG|\theta^{\Omega})$ holds. Hence, we formulate our first QM as follows:

$$\varphi_{\text{ssr}} = \frac{1}{\varphi_{\text{ef}}} \cdot -\frac{A}{N^{SG}}$$

$$A = SSR(\ell_2, g(\ell_1, \theta^{SG})) = \sum_{i=1}^{n^{SG}} \sum_{t=1}^{t_{\ell_1}^i} \left(\ell_{2t}^i - \hat{\alpha}^{SG}\ell_{1t}^i - \hat{\beta}^{SG}(\ell_{1t}^i - \hat{\psi}^{SG})_+\right)^2, \quad (7)$$

where $N^{SG} = \sum_{i=1}^{n^{SG}} t^i_{\ell_1}$ is the number of observations in the subgroup in target space and φ_{ef} is the entropy function [4] to discourage tiny subgroups. We take the SSR of ℓ_2 with respect to $g(\ell_1, \theta^{SG})$, which is defined in Equation (4). If the sum-of-squared error decreases, φ_{ssr} increases.

Although both the regression parameters and precision depend on the sum-of-squares, they are statistically independent. This means that we could find subgroups with a small error where $\ln p(SG|\theta^{SG}) > \ln p(SG|\theta^{\Omega})$ does not hold; the log likelihood of the subgroup may be large, but it may not be larger than the log likelihood of the global model, for instance when the regression parameters θ^{SG} do not differ much from θ^{Ω}. Therefore, we propose a QM that rewards not only small values of SSR for the subgroup, but also values of SSR for the subgroup that are smaller than the SSR of the subgroup evaluated on the global model:

$$\varphi_{\text{ssrb}} = \varphi_{\text{ef}} \cdot \frac{A(B-A)}{N^{SG}}, \tag{8}$$

where A is as in Equation (7) and B is similar but with θ^{SG} replaced by θ^{Ω}.

6 Experiments

We perform two experiments. First, we randomly sample 5% of the children and experiment with both QMs φ_{ssr} and φ_{ssrb}. We perform beam search [4, Algorithm 1] with $b = 4$, $w = 20$, and $q = 10$. Especially when working with domain-specific data, we aim for our resulting subgroup set to be a good balance between interpretability, variety, and quality. To further understand how a weighted coverage scheme (WCS) [14] can contribute to finding such a balanced subgroup set, and what its relation is to the search depth d, we vary $d \in \{3, 5\}$ and the multiplicative weighting parameter of the WCS $\gamma \in \{0.1, 0.5, 0.9\}$. We evaluate our results by inspecting the average quality of the subgroup set, the average size of the subgroups, the number of subgroups (out of $q = 10$) that validation with the Distribution of False Discoveries (DFD) [5] cannot distinguish from false discoveries over $m = 50$, the average run time, and two measures of subgroup set redundancy: Joint Entropy (JE) [14] and median Jaccard similarity (JSIM) [21] (see [25, Section 2] for precise definitions). We use the `pwlf` Python library to fit our segmented linear regression models [10].

Second, based on our findings in the first experiment, we choose the most appropriate QM, value for d and value for γ, and repeat the experiment with the full FUNA dataset ($n = 15\,486$). Beam search width $w = 20$, $b = 4$ and $q = 20$. All these children have at least 5% of their answers correct in each descriptor task (NC, SA, SS, CA) and the children have at least one observed answer for every possible set size in the DE task. The maximum number of observed items in the DE task is 18 per child. Our experimental code, all results, and a slice of the FUNA dataset are available at `https://github.com/RianneSchouten/FUNA_EMM`.

Extra experiments on Curran dataset We perform an additional set of experiments on a fully public dataset and find subgroups of children with exceptional relations between age and reading skills. Since our quality measures generalize to linear regression problems other than segmented linear regression, we perform these extra experiments with polynomial regression. More information and a short discussion of the results can be found in [25, Section 3].

6.1 Results Experiment 1

Figure 3 presents the standardized, average quality of a subgroup set ($q = 10$) for various values of d, γ, and both QMs. In essence, the results are as expected: the quality increases with the description length d and the weight parameter γ increases, and the impact of varying γ is larger for smaller d (see Figure 3; absolute difference between the smallest and largest quality for varying γ is larger for $d = 3$ than for $d = 5$). Table 4 reports the other evaluation metrics: the average subgroup size decreases when either d or γ increase, and in general, the subgroup set redundancy is larger when d decreases or γ increases (higher JE, lower JSIM). Except for 2 subgroups for φ_{ssrb} when $d = 3$ and $\gamma = 0.1$, all discovered subgroups can be considered valid discoveries.

For φ_{ssr}, given d, the average subgroup size, JE, and JSIM are comparable when varying values of γ. It seems that there is barely an effect of the WCS. When $d = 5$, the average quality is lower for $\gamma = 0.9$ then for $\gamma = 0.5$, and when $d = 3$, the average quality is lower for $\gamma = 0.5$ than for $\gamma = 0.1$. These results are unexpected since a decreasing γ is supposed to increase the variety

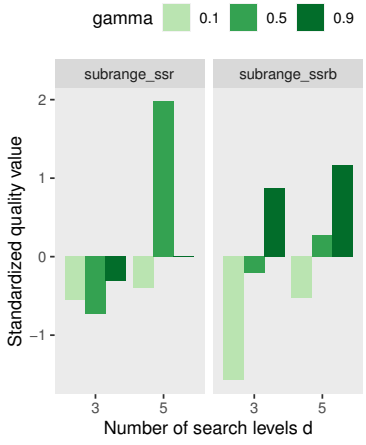

Fig. 3. The relation between the average quality of a subgroup set ($q = 10$, standardized per QM), search depth d, and WCS parameter γ, for both QMs.

Fig. 4. Experimental results for both QMs, $d \in \{3, 5\}$, $\gamma \in \{0.1, 0.5, 0.9\}$.

QM	d	γ	Prop	DFD	JE	JSIM	Time
φ_{ssr}	3	0.1	0.20	0	1.36	0.87	1.63
		0.5	0.16	0	1.68	0.75	1.61
		0.9	0.16	0	1.55	0.73	1.62
	5	0.1	0.12	0	0.91	0.88	2.97
		0.5	0.13	0	0.79	0.91	2.95
		0.9	0.13	0	0.77	0.90	2.93
φ_{ssrb}	3	0.1	0.22	2	4.35	0.18	1.60
		0.5	0.08	0	2.35	0.31	1.56
		0.9	0.05	0	1.14	0.44	1.35
	5	0.1	0.06	0	2.19	0.29	2.17
		0.5	0.06	0	2.01	0.31	2.13
		0.9	0.05	0	1.08	0.46	1.86

Table 3. Subgroup proportion, description and estimated target models for subgroups 1, 5, 6, 7, 10, 17 and 18, discovered with φ_{ssrb}. The global target model is $1407+88\ell_1+463(\ell_1-3.3)_+$

SG	Prop	Description	Target model
1	0.05	NC-IES:(0.04,1.0) $\wedge$ NC-MeanTC:(0.23,0.74) $\wedge$ SA-MeanTC:(0.71,1.0)	$2179+124\ell_1+764(\ell_1-2.9)_+$
5	0.50	NC-IES:(0.03,1.0)	$1541+106\ell_1+624(\ell_1-3.2)_+$
6	0.50	NC-IES:(0,0.03)	$1091+\ 70\ell_1+475(\ell_1-3.5)_+$
7	0.12	NC-MeanTC:(0.16,1.0) $\wedge$ SS-SumAnsC:(0.0,0.32)	$1938+\ 70\ell_1+712(\ell_1-2.8)_+$
10	0.38	SA-MeanTC:(0.71,1.0)	$1544+108\ell_1+641(\ell_1-3.2)_+$
17	0.30	grade = 3	$1561+106\ell_1+634(\ell_1-3.2)_+$
18	0.27	SA-MaxItem:(0.6,1.0)	$1064+\ 62\ell_1+441(\ell_1-3.6)_+$

of the subgroup set at the cost of average quality. Inspecting the individual descriptions and qualities, we find that for φ_{ssr} the variety in the subgroup set is larger when $\gamma = 0.9$ than when $\gamma \in \{0.1, 0.5\}$. Most likely, the reason is the use of a *square* when calculating the quality. Even when we use a strict WCS, the same subgroup recurs, since the weighted quality of the other subgroups does not beat the non-weighted quality of the recurring subgroup. When the WCS is very strict (small γ), at lower search levels, important precursors may be removed and not available for refinement at higher levels. As a consequence, a subgroup set with a strict WCS could have fewer candidate subgroups, which in the end creates a relatively redundant subgroup set. It is unfortunate that JE and JSIM do not fully reveal these conclusions. With φ_{ssrb} the subgroup sets are less redundant than with φ_{ssr}, especially for small values of γ. Clearly, JSIM increases and JE decreases when γ increases. Subgroups found with $d = 5$ are slightly smaller than for $d = 3$.

6.2 Results Experiment 1

We perform the experiment on the entire dataset with φ_{ssrb}, since this QM turns out to be stable and produces small and interesting subgroups. We choose $\gamma = 0.5$ to balance between high quality and low redundancy. We choose $d = 3$ since Table 4 shows that these results do not differ much from $d = 5$, and a description with fewer literals is easier to interpret for domain experts. Descriptions and target models of all top-20 exceptional subgroups can be found in [25, Section 2]; we report a smaller selection in Table 3 and Figure 5.

Although we allow for descriptions to have $d = 3$ literals, strong performance is found in single-attribute subgroups. There is a variety in used descriptors (multiple aggregation functions, multiple tasks), subgroup size, and target models. Compared to the segmented linear regression parameters of the global model, 15 out of 20 exceptional subgroups have a subitizing range lower than average; the other 5 have a higher subitizing range.

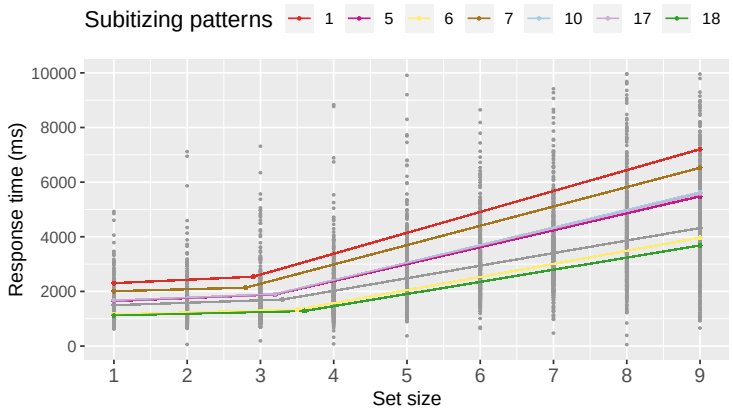

Fig. 5. Estimated segmented linear regression models of subgroups 1, 5, 6, 7, 10, 17 and 18 discovered with φ_{ssrb}. Target model equations can be found in Table 3.

Subgroups 1 and 2 have very similar subitizing curves: children in these subgroups are particularly slow to subitize, and these groups are the only ones that have an intercept over 2 seconds. The subgroups contain children with slow NC response times (either expressed in terms of IES or mean response time) and both are slow to solve addition problems (based on SA and CA tasks). The groups are small, and probably most typical of dyscalculia, or at the very least groups that are made up of children who are likely to have maths learning difficulties. The dyscalculia prevalence estimate is 3-6% [26], which is in accordance with the subgroup sizes 0.05 and 0.06 for subgroups 1 and 2 respectively.

Subgroup 5 is a more general version of subgroup 1; it covers 50% of the children and contains only the first literal. The subitizing curve shows the same trend as the one of subgroup 1, but less extreme: the subitizing range is smaller than the global model, but not as small as in subgroup 1, and intercept, subitizing slope, and counting slope are larger than in the global model, but not as large as in subgroup 1. Domain experts suspect that this subgroup may reflect maths learning difficulties as well.

Subgroup 6 is the inverse of subgroup 5. This is not only clear from the description in Table 3, but from the regression model in Figure 5 as well; the subitizing range is higher, and the intercept and subitizing slope are lower than in the global model. Subgroups 13, 15, 18, and 19 are the other four subgroups that have subitizing ranges above the average, and characteristically have subitizing intercepts (baseline response time or speed of processing) that are 300-350ms faster than the average and at least 500ms faster than any other group in the table. They also have shallower (faster) counting slopes by 150-200ms than most other groups.

Subgroups 18 and 19 have target models that are very similar to the one of subgroup 6, even though the descriptions of these subgroups differ. Subgroup 6 expresses the subgroup in terms of NC-IES while subgroup 18 does this in terms of an arithmetic addition task (SA). A similar things occurs for subgroups 5 and 10: the target models are similar while the descriptions use aggregated descriptors from different tasks. These findings suggests relations between number processing skills and arithmetic skills. They additionally show that it may be possible to obtain diagnostic information by focusing on fewer tasks; it may be possible to know the results on a particular task given the performance on another task. This is a promising result that provides opportunity for further development of assessment tools and intervention programs.

The only subgroup that does not use an aggregated descriptor is subgroup 17, which selects children in the third grade. Interestingly, the estimated target model of subgroup 17 is similar as the ones for subgroups 5 and 10; similar to the global model, expect for a larger counting slope. Compared to the other children in the FUNA dataset, the children in subgroup 17 are younger and hence, slower for all tasks, including the NC (subgroup 5) and SA (subgroup 10) tasks.

7 Discussion and Conclusion

The FUnctional Numerical Assessment (FUNA) project [22] develops digital assessment tools for detecting dyscalculia and dyslexia in young children by evaluating numerical processing competences such as the ability to enumerate small sets of dots and to compare the relative magnitudes between sets. These numerical processing competences are diagnostic markers of children's emerging math abilities [7]. In this paper, we particularly focus on the characteristics that define children's enumeration ability, such as the threshold at which children can determine the correct number of dots at a glance, known as subitizing range, and other parameters of subitizing patterns such as the initial reaction time and counting slope. Common algorithms used for estimating subitizing range can produce inconsistent results [15] especially among individuals with dot enumeration curves that deviate from the typical curve.

Therefore, we develop an EMM model class for segmented linear regression to discover subgroups of children whose subitizing curves exhibit atypical patterns. It could be argued that choosing segmented linear regression as a model class is a drawback since the observations are not independently distributed (i.e. a model is estimated on n^{SG} independent children, who all contribute the measurements of several items, resulting in a total number of N^{SG} observations). Despite of that, we follow this approach since segmented linear regression fits the neuro-cognitive concept of subitizing very well. Furthermore, the assumption of independent observations is required for most of the other algorithms as well; segmented linear regression has the least baggage built into it.

Our findings confirm the belief that numerical processing competences strongly correlate with arithmetic skills. We find several exceptional subgroups that confirm existing knowledge, including subgroups that are considered typical of dyscal-

culia; these children have slow NC response times and are slow to solve addition problems. We find subgroups with similar subitizing patterns but different descriptions. This indicates the strong relation between subitizing, counting, and arithmetic ability, and additionally provides promising opportunities for further development of assessment tools and intervention programs that focus on fewer tasks or a reduced number of items per task: it may become possible to know the results on a particular task given a child's performance on another task.

Both quality measures in this paper assume that the overall population and subgroups are best modelled with the canonical subitizing range model: a piecewise linear regression model with precisely one break point. However, it is entirely possible that coherent subgroups of children do not follow this regimen: some groups may display no substantial break point; behavior of others might be best modelled by multiple break points. The piecewise linear regression model class for EMM can accommodate this sort of behavior, but it requires development of a new QM: log likelihoods will necessarily increase when more break points are available to the model, so some penalty for model complexity must be involved.

Acknowledgments. This research is supported by the Exceptional and Deep Intelligent Coach (EDIC) project, partly funded by the Dutch Research Council (NWO). The tasks and data used in this project were designed and collected by the FUNA research consortium using the digital ViLLE learning platform offered by the University of Turku, Finland.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Benavides-Varela, S., Laurillard, D., Piperno, G., Fava Minor, D., Lucangeli, D., Butterworth, B.: Chapter 2 - digital games for learning basic arithmetic at home. In: Santos, F.H. (ed.) Game-Based Learning in Education and Health - Part A, Progress in Brain Research, vol. 276, pp. 35–61. Elsevier (2023)
2. Bishop, C.M.: Pattern recognition and machine learning. Springer (2006)
3. Bueno, M.L.P., Hommersom, A., Lucas, P.J.F.: Temporal exceptional model mining using dynamic bayesian networks. In: Proc. AALTD. pp. 97–112 (2020)
4. Duivesteijn, W., Feelders, A.J., Knobbe, A.: Exceptional Model Mining. Data Mining and Knowledge Discovery (DAMI) **30**(1), 47–98 (2016)
5. Duivesteijn, W., Knobbe, A.: Exploiting false discoveries: Statistical validation of patterns and quality measures in Subgroup Discovery. In: Proceedings of the IEEE International Conference on Data Mining (ICDM). pp. 151–160 (2011)
6. Feigenson, L., Dehaene, S., Spelke, E.: Core systems of number. Trends in cognitive sciences **8**(7), 307–314 (2004)
7. Gray, S.A., Reeve, R.A.: Preschoolers' dot enumeration abilities are markers of their arithmetic competence. PLoS One **9**(4), e94428 (2014)
8. Hellstrand, H., Holopainen, S., Korhonen, J., Räsänen, P., Hakkarainen, A., Laakso, M.J., Laine, A., Aunio, P.: Arithmetic fluency and number processing skills in identifying students with mathematical learning disabilities. PsyArXiv (2023). https://doi.org/10.31234/osf.io/jtk8c

9. Holloway, I.D., Ansari, D.: Mapping numerical magnitudes onto symbols: The numerical distance effect and individual differences in children's mathematics achievement. Journal of Experimental Child Psychology **103**(1), 17–29 (2009)
10. Jekel, C.F., Venter, G.: pwlf: A python library for fitting 1D continuous piecewise linear functions (2019), `https://github.com/cjekel/piecewise_linear_fit_py`
11. Krogel, M.A., Wrobel, S.: Transformation-based learning using multirelational aggregation. In: Proc. ILP. pp. 142–155 (2001)
12. Landerl, K., Bevan, A., Butterworth, B.: Developmental dyscalculia and basic numerical capacities: A study of 8–9-year-old students. Cognition **93**(2), 99–125 (2004)
13. Lavrač, N., Džeroski, S., Grobelnik, M.: Learning nonrecursive definitions of relations with LINUS. In: Proc. EWSL. pp. 265–281 (1991)
14. van Leeuwen, M., Knobbe, A.: Diverse Subgroup Set Discovery. Data Mining and Knowledge Discovery **25**(2), 208–242 (2012)
15. Leibovich-Raveh, T., Lewis, D.J., Kadhim, S.A.R., Ansari, D.: A new method for calculating individual subitizing ranges. Journal of Numerical Cognition **4**(2), 429–447 (2018)
16. Lemmerich, F., Becker, M., Singer, P., Helic, D., Hotho, A., Strohmaier, M.: Mining subgroups with exceptional transition behavior. In: Proc. KDD. pp. 965–974 (2016)
17. Major, C.S., Paul, J.M., Reeve, R.A.: TEMA and dot enumeration profiles predict mental addition problem solving speed longitudinally. Frontiers in psychology **8**, 313803 (2017)
18. Muggeo, V.M.: Estimating regression models with unknown break-points. Statistics in Medicine **22**(19), 3055–3071 (2003)
19. Muggleton, S.: Inverse entailment and Progol. New generation computing **13**, 245–286 (1995)
20. Mulders, P.J.A.M., van den Heuvel, E.R., Reidsma, P., Duivesteijn, W.: Introducing exceptional growth mining—analyzing the impact of soil characteristics on on-farm crop growth and yield variability. PLOS ONE **19**(1), 1–26 (01 2024)
21. Proença, H.M., Grünwald, P., Bäck, T., van Leeuwen, M.: Discovering outstanding subgroup lists for numeric targets using MDL. In: Proc. ECML PKDD (2020)
22. Räsänen, P., Aunio, P., Laine, A., Hakkarainen, A., Väisänen, E., Finell, J., Rajala, T., Laakso, M.J., Korhonen, J.: Effects of gender on basic numerical and arithmetic skills: Pilot data from third to ninth grade for a large-scale online dyscalculia screener. Frontiers in Education: Educational Psychology **6**, 683672 (2021)
23. Reeve, R., Reynolds, F., Humberstone, J., Butterworth, B.: Stability and change in markers of core numerical competencies. Journal of Experimental Psychology: General **141**(4), 649 (2012)
24. Schouten, R.M., Bueno, M.L., Duivesteijn, W., Pechenizkiy, M.: Mining sequences with exceptional transition behaviour of varying order using quality measures based on information-theoretic scoring functions. Data Mining and Knowledge Discovery **36**, 379–413 (2022)
25. Schouten, R.M., Duivesteijn, W., Räsänen, P., Paul, J.M., Pechenizkiy, M.: Exceptional Subitizing Patterns — Supplementary Material. Tech. rep., available at Figshare, `https://doi.org/10.6084/m9.figshare.26008879` (2024)
26. Shalev, R.S., Auerbach, J., Manor, O., Gross-Tsur, V.: Developmental dyscalculia: prevalence and prognosis. European child & adolescent psychiatry **9**, S58–S64 (2000)
27. Song, H.: Model-based Subgroup Discovery. Ph.D. thesis, Un. of Bristol (2017)
28. Železný, F., Lavrač, N.: Propositionalization-based relational subgroup discovery with RSD. Machine Learning **62**, 33–63 (2006)