Stealth Assessment in ITS - A Study for Developmental Dyscalculia

Severin Klingler[1], Tanja Käser[1], Alberto-Giovanni Busetto[2], Barbara Solenthaler[1], Juliane Kohn[3], Michael von Aster[3,4,5], and Markus Gross[1]

[1] Department of Computer Science, ETH Zurich, Switzerland
[2] Department of Electrical and Computer Engineering, University of California, USA
[3] Department of Psychology, University of Potsdam, Germany
[4] Center for MR-Research, University Children's Hospital Zurich, Switzerland
[5] Department of Child Adolescent Psychiatry, DRK Kliniken Berlin Westend, Germany

Abstract. Intelligent tutoring systems are adapting the curriculum to the needs of the student. The integration of stealth assessments of student traits into tutoring systems, i.e. the automatic detection of student characteristics has the potential to refine this adaptation. We present a pipeline for integrating automatic assessment seamlessly into a tutoring system and apply the method to the case of developmental dyscalculia (DD). The proposed classifier is based on user inputs only, allowing non-intrusive and unsupervised, universal screening of children. We demonstrate that interaction logs provide enough information to identify children at risk of DD with high accuracy and validity and reliability comparable to traditional assessments. Our model is able to adapt the duration of the screening test to the individual child and can classify a child at risk of DD with an accuracy of 91% after 11 minutes on average.

Keywords: automatic assessment, feature processing, Bayesian network, pairwise clustering, computer-based screening, dyscalculia

Intelligent tutoring systems (ITS) are gaining importance in education. A lot of research has been conducted to represent and model student knowledge accurately, design effective curricula and develop optimal instructional policies. A large body of work has focused on mining the data logs collected from ITS. Important topics in this area are automatic stealth assessments such as the evaluation of student learning or detection of student properties (e.g. intelligence, learning disabilities) [31]. Traditional assessments are often time consuming and have to be supervised by an expert, rendering them expensive in practice. Hence, this approach does not scale and is therefore not suitable in many cases, such as MOOCs, large university courses, or widespread screenings in elementary schools to enable early detection of learning disabilities.

Previous work has investigated stand-alone automatic digital assessments, including research on automatic scoring [5], item generation [18] and game-based assessment [20]. Furthermore, digital screening programs replacing traditional neuropsychological tests, for example for dyscalculia [10] or dyslexia [12], have been developed. Ideally, such computer-based screening programs are seamlessly

integrated into an ITS. This enables not only automatic and non-intrusive assessment of students, but also analysis and detection of student traits that allow for a better adaptation of the curriculum to the individual needs. Despite these advantages only few work have addressed such ITS systems with fully integrated assessment. One step in this direction are integrated behavior detectors identifying students gaming the system [6], finding wheel-spinning students [9] or modeling engagement, e.g. [8, 1, 14]. Other work used clustering and classification approaches to detect students' mathematical characteristics [23].

In this paper, we propose a pipeline for integrating automatic assessment, i.e. detectors of student traits, directly into the tutoring system. We validate our approach for the case of developmental dyscalculia (DD) (a specific learning disability affecting the acquisition of arithmetic skills [2]) and the game-based training environment *Calcularis* [22].

Our pipeline leverages the potential of machine learning algorithms. Its data-driven nature features several advantages. First, since it builds upon a large set of student training data, the costs for model building are low and the accuracy of the classifier can be continuously improved as more student data is added over time. The test duration can be adapted to each child individually, which reduces the average test duration substantially. Second, our classifier can be seamlessly embedded into an ITS (in our case *Calcularis* [22]), where the assessment runs continuously and non-intrusively in the background. This integration reduces testing expenses and emotional stress imposed to children is kept at a minimum. The embedding allows the ITS to leverage the information from the stealth assessment during the training. Third, our pipeline has the potential to be applied to a different ITS and be used for the assessment of different student traits.

We extensively evaluate the accuracy, practicability, and validity of our approach on data logs from 68 children. Our results demonstrate that we can identify children at risk of DD with a high accuracy (91% sensitivity, 91% specificity) within a short time (11 minutes on average). We conclude from our results that recorded user inputs alone could potentially allow for a detailed reconstruction of student traits and that the integration of stealth assessments may refine the adaptation of the curriculum that ITS are currently providing.

1 Adaptive Classification Algorithm

Our adaptive classification is based on the training environment *Calcularis* [22], a computer-based system for learning mathematics designed for children with DD. The program is structured into different instructional games, which are designed based on current neuro-cognitive theory. *Calcularis* consists of ten different games representing 100 different skills that are essential for learning mathematics. Our model building process consists of four steps (see Figure 1). We first extract a large set of candidate features and then perform feature selection based on common similarity measures. Next, we build our adaptive classifier by first sorting the selected features and then defining a Naive Bayes model.

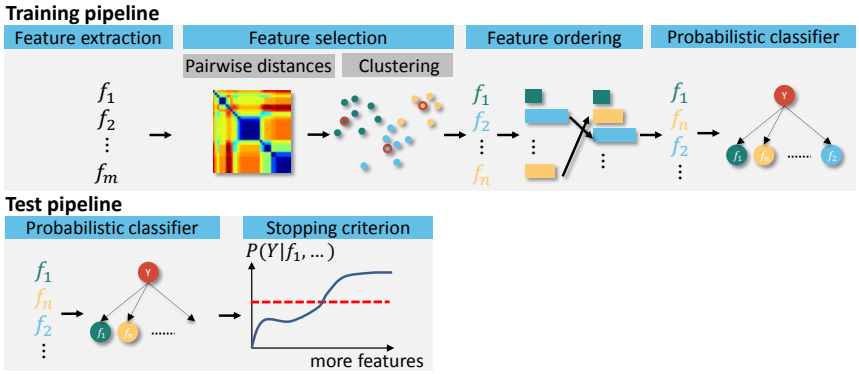

Fig. 1. Processing pipeline: Pairwise distances of features f serve as input for the clustering. We select the representative feature per cluster and determine an optimal feature ordering. A Naive Bayes model is trained on the selected features. The probabilistic output of the classifier is used to adapt the test duration to each child.

Feature extraction. We identified a set of recorded features that describe different mathematical properties of the user. These features can be classified into *skill-* and *game dependent* features, and are summarized in Table 1. *Skill dependent* features provide information about tasks associated with a specific skill. The performance **P** for a skill measures the ratio of correctly solved tasks for a given number of tasks. We expect children without DD to outperform children with DD on these tasks, since mathematical abilities of children with DD are at a level comparable to the level of children without DD of lower age [3]. Answer time **AT** is measured for all skills as children with DD tend to have longer answer times compared to children without DD [17]. They often show deficits in fact retrieval and tend to have difficulties to acquire arithmetic procedures [28] which increases answer times for simple arithmetic tasks. We count typical mistakes **TM** for a subset of games where such a measure is meaningful. **TM** are extracted by matching the erroneous result to a set of error patterns. As an example switching the digits of the result in an arithmetic task is considered a typical mistake (e.g. $15 + 9 = 42$). The complete set of error patterns is described in [22]. Additional *game dependent* features were chosen related to specific games. The estimation game feature **E** measures the relative number of overestimates when estimating the number of points in a point cloud. Whether children with DD are less sensitive to differences in this number representation is not consistently supported by recent work [27]. The feature **SN** for the secret number game measures the reduction of the search interval while repeatedly guessing the same number. This feature quantifies common problem-solving strategies such as bisection of the search interval or linear search. The ordering game feature **O** measures the ratio of false positives when assessing whether numbers are in ascending order. Children with DD are shown to be less efficient when processing numbers [26], therefore we hypothesize that they will perform worse when comparing numbers. The landing game feature **L** measures the error of the number estimate. Deficits in spatial number representation as often shown

Feature	Description
Skill dependent features (extracted at specific skills)	
Performance	Ratio of correctly solved tasks.
Answer **T**ime	Average answer time.
Typical **M**istakes	Number of typical mistakes committed.
Game dependent features	
Estimation	Estimating the number of displayed points.
	E is the ratio between number of overestimates and task count.
Secret **N**umber	Guessing a number in as few steps as possible.
	S is the ratio by which the remaining search interval is reduced.
Ordering	Is a number sequence ordered ascending?
	O is the ratio of false positive and incorrectly solved tasks.
Landing	Positioning a number on a number line.
	L is the distance to the correct position of the given number).

Table 1. Extracted features and abbreviations (bold) used in the screener.

by children with DD [25] are obstructive to this task, thus we expect children with DD to perform significantly worse compared to peers without DD.

Feature selection. Our feature extraction yields a few hundred features, each corresponding to a set of tasks the user has to solve. Therefore, the number of features directly influences the test duration. To limit the test duration and to remove possible correlations between features, we only use a subset of features for classification. We cluster the features into groups based on their similarity and select one representative feature per cluster. As the different feature types have different domains (e.g., $\mathbf{P} \in [0,1]$, **AT** seconds > 0) a direct comparison between the features is not meaningful. We therefore process the features to make them comparable.

In a first step, we compute a similarity matrix $\mathbf{K_i} \in [0,1]^{S \times S}$ for each feature f_i, where S denotes the number of children. Therefore, $\mathbf{K_i}$ contains the pairwise similarities between each pair of children regarding feature f_i. We design the matrices based on the nature of each feature and in particular exploiting invariance of the feature types. For example, for the answer time **AT** we combine a Gaussian kernel with a log transform to obtain

$$\mathbf{K_i}(s,u) = exp\left(-\frac{\|log(f_i^s) - log(f_i^u)\|^2}{2\sigma^2}\right), \tag{1}$$

where f_i^s and f_i^u denote the respective feature values for children s and u. We incorporate a cumulative beta distribution to design the similarity matrix for the performance features **P**. For the **SN** feature, we designed an exponential kernel. All other features (**TM**,**E**,**O**,**L**) apply a standard Gaussian kernel. Further details regarding the design of the different kernels can be found in [24].

In a second step, we cluster the features using pairwise-clustering [21] based on the pairwise distances $d_{ij} = \|\mathbf{K_i} - \mathbf{K_j}\|_F$ between all feature pairs using the Frobenius norm. We then compute an optimal matrix $\mathbf{T}$, which contains the pairwise Hamming distances between child labels, i.e., $\mathbf{T}(s,u) = 0$ if s and u belong to the same group $\in \{DD, CC\}$, with CC referring to control, and

$\mathbf{T}(s,u) = 1$ otherwise. For each cluster, we select one representative feature, which is the one with the smallest distance $dt_i = ||\mathbf{K_i} - \mathbf{T}||_F$ to matrix $\mathbf{T}$.

Probabilistic classifier. Based on the selected features, we develop a probabilistic model that adapts the test duration to the individual child. The classification task is solved using an adapted Naive Bayes model, which assumes conditional independence of all the features f_i given the group label Y ($Y = 0$ child with DD, $Y = 1$ CC), but was shown to perform optimally even if the independence assumption is violated [34]. Correlations between features are low in our case (average ρ=0.07, <1% significant correlations at α=0.001) because of our feature selection step. The posterior probability of the group label Y for a child given N observed features is proportional to

$$p(Y|f_1, ..., f_N) \propto \prod_{i=1}^{N} p(f_i|Y) \cdot p(Y), \tag{2}$$

where for every feature we choose the density $p(f_i|Y)$ from a set of standard distributions that best models the data according to the BIC score. We assume a normal distribution for the features **E**, **SN**, **O** and **L**, and a Beta, Gamma, and Poisson distribution for **P**, **AT**, and **TM**, respectively. The prior probability $p(Y)$ is set to the estimated prevalence of DD [30]. Due to the independence assumption, we can deal with cases where we only observe a subset of all features. After observing the first feature f_1, we can compute $p(Y = 1|f_1)$. Having observed f_2, we infer $p(Y = 1|f_1, f_2)$ etc. For any threshold $\tau \in [0, 1]$, the predicted group label $\hat{Y}$ can then be computed as

$$\hat{Y} = \begin{cases} 1 & p(Y = 1|f_1, ..., f_n) > \tau \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Feature ordering. To determine the optimal ordering of the tasks in the test, we compute the amount of group information contained in each feature. We prefer features where the feature values differ substantially across the groups (DD and CC) and are similar within the group. To assess the quality of each feature f_i, we use an unpaired t-test for a difference in means of the two independent groups. We then order the features by sorting the calculated p-values in ascending order, *i.e.*, the feature with the smallest p-value is asked first.

Stopping criterion. The optimal point in time to stop the test is heuristically determined. After observing the first t features the classifier has a current belief about the group label of a child and predicts the label based on $p(Y|f_1, \ldots, f_t) > \tau$ (see Equation (2)). Intuitively, we stop the test if observing the next feature would not contradict our current belief about the group label. As the next feature value f_{t+1} is unknown, the feature value in the training data $\hat{f}_{t+1}$ that contradicts the model's current belief the most is taken instead. We stop if observing $\hat{f}_{t+1}$ is not changing the current belief, i.e., if $p(Y = 1|f_1, ..., f_t) > \tau$ and $p(Y = 1|f_1, ..., f_t, \hat{f}_{t+1}) > \frac{\tau}{2}$.

Type/Nr.	Order	Skill
AT/1	2	Addition 2,1 TC* ('13+8=21')
AT/2	14	Point set estimation
AT/3	4	Subtraction 3,1 TC* ('122-7=115')
AT/4	8	Addition 3,1 TC* ('128+4=132')
AT/5	15	Are numbers sorted ascending
AT/6	12	Spoken to written number
P/1	5	Spoken to written number
P/2	10	Subtraction 2,2 TC* ('56-38=18')
P/3	1	Find neighbor numbers ±10
P/4	11	Spoken to written number
SN/1	7	Guess a number
SN/2	13	Guess a number
TM/1	3	Subtraction 2,1 TC* ('74-9=65')
TM/2	9	Assign spoken number to number line
TM/3	16	Addition 2,2 TC* ('23+18=41')
TM/4	6	Subtraction 2,2 ('48-36=12')
TM/5	17	Assign written number to number line

* TC : with carrying / borrowing

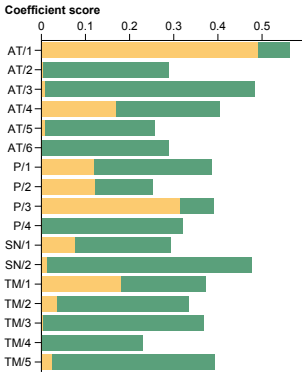

Fig. 2. Selected features and their corresponding skills and ordering in the test. The relationship between a feature and the test score is shown on the right, using Pearson's correlation coefficient (yellow) and the maximal information coefficient MIC (green).

2 Experimental Evaluation

The experimental evaluation of our method was based on log files from 68 participants (32 DD, 36 CC) of a multi-center user study conducted in Germany and Switzerland [32]. During the study, children trained with *Calcularis* at home for five times per week during six weeks and solved on average 1551 tasks. There were 28 participants in the 2^{nd} grade (9 DD, 19 CC) and 40 children in the 3^{rd} grade (23 DD, 17 CC). The diagnosis of DD was based on standardized neuropsychological tests [4, 19, 16].

We calculated the accuracy, the specificity and the sensitivity of our model based on the predicted and the true label of the students (either DD or CC). All results were computed on unseen students in the test set. Training and test sets were created using .632 bootstrap with resampling ($B = 300$). All parameter estimates are based on maximum likelihood estimation using Nelder-Mead simplex direct search. The optimization stops when the improvement in the likelihood is $< 10^{-4}$ or after 400 iterations. Hyper parameters (parameters for kernels and features) and features (including feature ordering) were selected using nested cross validation, employing .632 bootstrap with resampling ($B = 300$) on top of 10-fold cross validation. The optimal number k^* of clusters in the feature selection step was heuristically determined by limiting the maximal test duration to <35 minutes. Since we required five recorded tasks per feature (average recorded task time: 0.39 minutes), this test duration results in $k^* = 17$ clusters (which leads to 85 tasks in the test).

Content validity. 17 features were automatically selected based on the recorded data alone. For all features we calculated Pearson's correlation coefficient ρ^2 and the maximal information coefficient (MIC) [29] between the feature and the test score to measure the linear and non-linear relationships, respectively. For most features the relationship is highly non-linear, which prohibits the use of simple

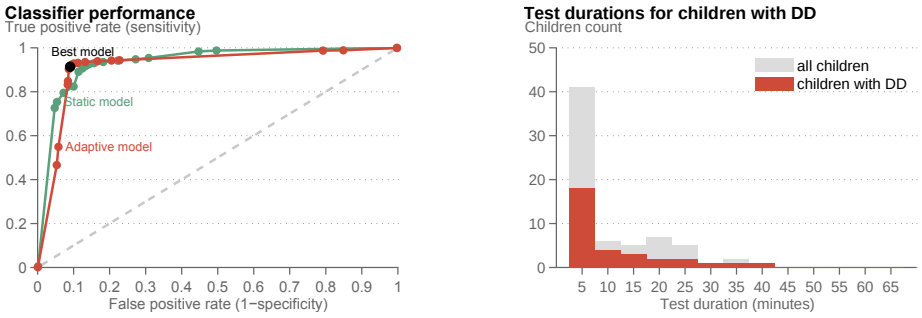

Fig. 3. Left. Performance comparison of the classifiers using ROC curves. The adaptive approach with reduced test duration (red) shows comparable performance to the classifier using all features (green). Points on the curves correspond to different probability thresholds τ at which the model decides if a child has DD. **Right.** Test durations for all children (grey) and DD (red). Our adaptive screener requires on average 11 test minutes to classify a child. Around 40% can be classified already after 5 minutes.

prediction methods such as linear regression. The feature ordering yields the optimal task sequence in the test as listed in Figure 2.

The automatically selected features agree well with findings in previous work on DD. Deficits in number comparison that are shown by children with DD [26] are captured by considering temporal and performance values (AT/5, P/3). Children with DD exhibit deficits in number processing [13]. Number processing skills are captured in various features and include again temporal and performance information (AT/2, AT/6, P/1, P/4). The features extracted from the number line game (TM/2, TM/5) capture typical mistakes in spatial number representation [11]. Furthermore, different problem solving strategies are analyzed based on the Secret Number game (SN/1, SN/2). Finally, difficulties acquiring simple arithmetic procedures and deficits in fact retrieval that are frequently shown by children with DD [28] are captured measuring answer times for various arithmetic procedures in AT/1, AT/3. Interestingly, no features from tasks associated with subitizing are selected, although subitizing is considered one of the basic functions often impaired for children with DD [26]. Most of the selected features correspond well with the type of tasks used in standardized tests for DD such as counting, number comparison, number representation and simple arithmetical tasks [4]. Note that the screener includes some features such as typical mistakes and problem solving strategies that are not captured by paper tests. The type of the selected features agrees well with other screening tools that measure answer time, performance and typical mistakes on tasks such as dot enumeration, number comparison, single digit arithmetic (Dyscalculia Screener Digital [10]) or recognizing reading and writing of natural number (DyscaliUM [7]).

Criterion-related validity. In Figure 3, left, we compare the performance of the static and adaptive Bayesian network model with ROC curves. In the static case (green line), we used all features, *i.e.*, all tasks, while in the adaptive case (red line) we used early test abortion based on our stopping criterion. Every point on the curves corresponds to a different threshold τ for the probabilistic classifier.

Our best classifier (selected by cross validation) exhibits a high sensitivity and specificity of 0.91 for a threshold $\tau = 0.3$ (black dot).

There is no significant decrease in performance when we stop the test early with our adaptive model, i.e., on average, children are not misclassified more frequently. In fact, the adaptive classifier that is based on partial data is outperforming the static approach for a specificity in the range $[0.05, 0.15]$. As the features are ordered based on how much information they carry about the group label, it can be advantageous to neglect those with little information since they tend to have more noisy information. Our classifier achieves a higher sensitivity compared to the stand-alone digital screening test DyscalculiUM; no comparison can be done with the Dyscalculia Screener Digital as it was standardized independent of traditional tests for DD.

Construct validity. Construct validity of our method was assessed by correlating the probabilistic output of our screener with a series of tests measuring different cognitive aspects of all participants. We performed standardized tests to asses convergent validity and discriminant validity as listed on the right. We observe moderate to high correlation coefficients for all measures capturing related cognitive concepts and weak correlations to the set of tests measuring unrelated concepts. These results are comparable to construct validity analysis of standardized neuropsychological tests that assess mathematical abilities. Correlations for these tests range from 0.22 to 0.73 [33, 15].

Test	ρ	p-value
Convergent validity		
Non verbal intelligence [16]	0.44	$<10^{-3}$
Math anxiety test [26]	0.42	$<10^{-2}$
Cognitive competence [1]	0.63	$<10^{-7}$
Discriminant validity		
Working memory [19]	0.19	0.13
Verbal intelligence [16]	0.23	0.06
Sport competence [1]	-0.17	0.18
Peer acceptance [1]	0.08	0.51
Attentional performance [43]	0.25	0.10

Reliability. Classical notion of test reliability in terms of measures such as Cronbach's alpha do not apply for our adaptive test due to non tau-equivalence of the measurements and the fact that our test output is a non-linear function of item scores. We therefore investigate the split-half reliability of our proposed model as an approximation to the standard notion of test reliability. We observe a reliability of 0.87. This is comparable to other mathematical tests where a reliability in the range of 0.7 to 0.92 is reported [15, 16].

Test duration. Due to our stopping criterion, the test duration is adapted to the individual child. Figure 3, right, shows the test duration for all children (grey) and for DD (red). On average, our adaptive screener classifies a child as DD or CC after only 11 minutes (at which point the test is stopped). This is notably shorter than screener durations reported in previous work. In comparison, the test duration of the Dyscalculia Screener Digital is reported to be between 15 and 30 minutes [10]. For Higher Education, a test duration of 48 minutes was reported using the computer-based screener for DD DyscalculiUM. With our adaptive screener, roughly 40% of children are already classified after five test minutes. Our static screener test takes 26.6 minutes on average, which emphasizes the importance of the adaptivity. The adaptive stopping criterion is important to retain classification accuracy as for 43% of the children the initial classification changed until the stopping criterion was met.

3 Discussion & Conclusion

We developed a fully data-driven pipeline for the automatic detection of student traits that can be seamlessly embedded into an ITS. We validated the method for the case of DD, allowing for non-intrusive and unsupervised screening of children while they are training with the ITS. The automatically selected features are covering a broad range of different characteristics of the children and are in accordance with the literature on DD. The classifier exhibits high sensitivity (0.91) and specificity (0.91) and adapts the test duration to each child individually, resulting in an average duration of as little as 11 minutes. Further, our method exhibits good construct validity (high correlations to tests measuring mathematical abilities, low correlations to tests assessing dissimilar abilities). These findings demonstrate that student traits can be effectively learned from user inputs alone. This knowledge about student traits allows an ITS to further adapt the curriculum to the specific needs of the students. In the future we would like to investigate potential intervention strategies based on the inferred knowledge about student traits. While this work evaluates the proposed model only for the screening of children at risk of DD, there is nothing inherently DD specific in the method. As such, our framework can be applied for the unobtrusive detection of other student traits and using different learning environments.

Acknowledgments. This work was supported by ETH Grant ETH-23 13-2.

References

1. Arroyo, I., Woolf, B.P.: Inferring learning and attitudes from a Bayesian Network of log file data. In: Proc. AIED. pp. 33–40 (2005)
2. von Aster, M.G., Shalev, R.: Number development and developmental dyscalculia. Developmental Medicine and Child Neurology 49, 868–873 (2007)
3. von Aster, M.: Developmental cognitive neuropsychology of number processing and calculation: varieties of developmental dyscalculia. Eur. Child&Adol. Psych. (2000)
4. von Aster, M., Zulauf, M.W., Horn, R.: Neuropsychologische Testbatterie für Zahlenverarbeitung und Rechnen bei Kindern: ZAREKI-R. Pearson (2006)
5. Attali, Y.: Reliability-Based Feature Weighting for Automated Essay Scoring. Applied Psychological Measurement 39(4), 303–313 (2015)
6. Baker, R.S., Corbett, A.T., Koedinger, K.R.: Detecting Student Misuse of Intelligent Tutoring Systems. In: Proc. ITS. pp. 531–540 (2004)
7. Beacham, N., Trott, C.: Screening for dyscalculia within HE. MSOR 5, 1–4 (2005)
8. Beck, J.E.: Engagement tracing: Using response times to model student disengagement. In: Proc. AIED. pp. 88–95 (2005)
9. Beck, J.E., Gong, Y.: Wheel-spinning: Students who fail to master a skill. In: Proc. AIED. pp. 431–440 (2013)
10. Butterworth, B.: Dyscalculia screener. Nelson Publishing Company Ltd. (2003)
11. Butterworth, B., Varma, S., Laurillard, D.: Dyscalculia: From brain to education. Science 332(6033), 1049–1053 (2011)
12. Cisero, C., Royer, J., Marchant, H., Jackson, S.: Can the computer-based academic assessment system (CAAS) be used to diagnose reading disability in college students? Journal of Educational Psychology 89(4), 599–620 (1997)
13. Cohen Kadosh, R., Cohen Kadosh, K., Schuhmann, T., Kaas, A., Goebel, R., Henik, A., Sack, A.T.: Virtual dyscalculia induced by parietal-lobe TMs impairs automatic magnitude processing. Current Biology 17, 689–693 (2007)

14. Cooper, D., Muldner, K., Arroyo, I., Woolf, B., Burleson, W.: Ranking Feature Sets for Emotion Models Used in Classroom Based Intelligent Tutoring Systems. In: UMAP, vol. 6075, pp. 135–146 (2010)
15. Desoete, A., Grégoire, J.: Numerical competence in young children and in children with mathematics learning disabilities. Learn. Individ. Differ. 16(4), 351–367 (2006)
16. Esser, G., Wyschkon, A., Ballaschk, K.: BUEGA: Basisdiagnostik Umschriebener Entwicklungsstörungen im Grundschulalter. Hogrefe, Göttingen (2008)
17. Geary, D.C., Brown, S.C., Samaranayake, V.A.: Cognitive addition: A short longitudinal study of strategy choice and speed-of-processing differences in normal and mathematically disabled children. Dev. Psychol. 27(5), 787–797 (1991)
18. Graf, E.A., Fife, J.H.: Difficulty Modeling and Automatic Generation of Quantitative Items: Recent Advances and Possible Next Steps. In: Automatic Item Generation: Theory and Practice, pp. 157–179. Routledge (2013)
19. Haffner, J., Baro, K., Parzer, P., Resch, F.: Heidelberger Rechentest (HRT): Erfassung mathematischer Basiskomptenzen im Grundschulalter (2005)
20. Hao, J., Shu, Z., Davier, A.: Analyzing Process Data from Game/Scenario- Based Tasks: An Edit Distance Approach. JEDM 7 (2015)
21. Hofmann, T., Buhmann, J.M.: Pairwise data clustering by deterministic annealing. IEEE Trans. Pattern Anal. Mach. Intell. 19(1), 1–14 (1997)
22. Käser, T., Baschera, G.M., Kohn, J., Kucian, K., Richtmann, V., Grond, U., Gross, M., von Aster, M.: Design and evaluation of the computer-based training program Calcularis for enhancing numerical cognition. Front. in Dev. Psychol. 4(489) (2013)
23. Käser, T., Busetto, A.G., Solenthaler, B., Kohn, J., von Aster, M., Gross, M.: Cluster-based prediction of mathematical learning patterns. In: Proc. AIED (2013)
24. Käser, T.: Modeling and Optimizing Computer-Assisted Mathematics Learning in Children. Ph.D. thesis, Diss., ETH Zürich, Nr. 22145 (2014)
25. Kucian, K., Grond, U., Rotzer, S., Henzi, B., Schönmann, C., Plangger, F., Gälli, M., Martin, E., von Aster, M.: Mental Number Line Training in Children with Developmental Dyscalculia. NeuroImage 57(3), 782–795 (2011)
26. Landerl, K., Bevan, A., Butterworth, B.: Developmental dyscalculia and basic numerical capacities: a study of 8-9-year-old students. Cognition 93, 99–125 (2004)
27. Noël, M.P., Rousselle, L.: Developmental changes in the profiles of dyscalculia: an explanation based on a double exact-and-approximate number representation model. Frontiers in Human Neuroscience 5, 165 (2011)
28. Ostad, S.A.: Developmental differences in addition strategies: A comparison of mathematically disabled and mathematically normal children. British Journal of Education Psychology 67, 345–357 (1997)
29. Reshef, D.N., Reshef, Y.A., Finucane, H.K., Grossman, S.R., McVean, G., Turnbaugh, P.J., Lander, E.S., Mitzenmacher, M., Sabeti, P.C.: Detecting novel associations in large data sets. Science 334(6062), 1518–1524 (2011)
30. Shalev, R., von Aster, M.G.: Identification, classification, and prevalence of developmental dyscalculia. Enc. of Language and Literacy Development pp. 1–9 (2008)
31. Shute, V.J.: Stealth assessment in computer-based games to support learning. Computer games and instruction (2011)
32. Von Aster, M., Rauscher, L., Kucian, K., Käser, T., McCaskey, U., Kohn, J.: Calcularis - Evaluation of a computer-based learning program for enhancing numerical cognition for children with developmental dyscalculia (2015), 62nd Annual Meeting of the American Academy of Child and Adolescent Psychiatry
33. Woolger, C.: Wechsler Intelligence Scale for Children-Third Edition (WISC-III). In: Understanding Psychological Assessment, pp. 219–233 (2001)
34. Zhang, H.: The Optimality of Naive Bayes. In: Proc. FLAIRS (2004)