DIFFERENCES IN QUANTITY RECOGNITION OF STUDENTS WITH AND WITHOUT MATHEMATICAL DIFFICULTIES: ANALYSIS THROUGH EYE-TRACKING AND AI

Maike Schindler[1], Erik Schaffernicht[2], Achim J. Lilienthal[2]

[1]University of Cologne, [2]Örebro University

Difficulties in mathematics learning are an important topic in practice and research. In particular, researchers and practitioners need to identify students' needs for support to teach and help them adequately. However, empirical research about group differences of students with and without mathematical difficulties (MD) is still scarce. Previous research suggests that students with MD may differ in their quantity recognition strategies in structured whole number representations from students without MD. This study uses eye-tracking (ET), combined with Artificial Intelligence (AI), in particular pattern recognition methods, to analyze group differences in gaze patterns in quantity recognition of N=164 fifth grade students.

INTRODUCTION

Learning difficulties in mathematics are an important topic in practice and research and have attracted increased interest not least since inclusive education has gained significance. Researchers and practitioners aim to understand knowledge and learning in a fine-grained way, and to foster students with MD individually and adequately (e.g., Moser Opitz et al., 2016; Scherer et al., 2016).

Previous research has indicated that students' strategies in quantity recognition can be used to identify difficulties in mathematics learning (e.g., Schleifer & Landerl, 2011; Schindler et al., 2019). An important question is how students' quantity recognition strategies can be observed. A promising method for investigating students' strategies in whole number representations such as the abacus or dot field is analyzing students' eye movements through ET (Lindmeier & Heinze, 2016; Rottmann & Schipper, 2002). Qualitative eye movement analyses in such representations may even outperform thinking aloud analyses in precision and level of detail—especially for students with MD (Schindler & Lilienthal, 2018). However, the qualitative analysis of eye movements is laborious and potentially subjective, and it is not yet sufficiently clear how eye movements in quantity recognition tasks differ between students with and without MD. Even though initial studies report group differences in students' strategy use (Schindler et al., 2019), the statistical analyses and the explanatory power of the results are limited due to small sample sizes of students. Therefore, this study aims to investigate group differences in students' gaze patterns when determining quantities in structured whole number representations—with a computer-supported, i.e., automated, evaluation method. We pursue the research question *Do the gaze patterns of MD vs. non-MD students in quantity recognition in structured whole number representations differ?,* which contributes to our overall purpose to investigate whether we can

2019. In M. Graven, H. Venkat, A. Essien & P. Vale (Eds.). *Proceedings of the 43rd Conference of the International Group for the Psychology of Mathematics Education* (Vol. 3, pp 281-288). Pretoria, South Africa: PME.

automatize the analysis of group differences in ET data and the evaluation to which group a student belongs.

Automatization of the analysis of ET data is desirable because analyzing eye movements for identifying student strategies is extraordinary time-consuming and demanding, given that there are—depending of the framerate of the device—100 or even 1000 frames per second recorded by the eye-tracker, all of which need to be analyzed if strategies are to be identified in videos (like in Schindler & Lilienthal, 2018, 2019). The effort required to manually analyze ET data is also prohibitive if ET based methods are to be used routinely by practitioners, e.g., school teachers.

Motivated by these considerations, we apply a methodology that makes use of a set of pattern recognition methods from AI. In order to compare eye gazes of the groups of MD vs. TD (typically developing) students, we analyze differences in students' gaze patterns on digital task sheets. In general, gaze patterns between the groups of students could differ in a myriad of ways that arise from combinations of where, when, for how long and in which sequence the students look at the stimuli. To render the subsequent analysis feasible, we first select a reduced representation: heat maps, i.e., visual representations displaying all gazes for each task. Simply put, we then investigate whether the heat maps of the two groups can be separated well on task level by our pattern recognition system, indicating significantly different gaze patterns. Our approach even allows us to semantically interpret group differences: The analysis of group-averaged heat maps (displaying all MD (vs. TD) students' gazes for the tasks, Fig. 2) allows us to identify differences that are meaningful for mathematics education research and hint at how strategy use might be different between the groups.

MATHEMATICAL DIFFICULTIES

To date there is no common definition or term describing the group of students having difficulties in mathematics (Scherer et al., 2016). Terms such as *mathematical learning disabilities*, (severe) *mathematical difficulties*, or *developmental dyscalculia* are used—depending on different educational contexts and research traditions. Medical models label a *disorder* (e.g., WHO, 2018) and support an IQ-discrepancy model. However, recent research suggests not to distinguish between students with MD depending on the discrepancy between their IQ and their math performance, since cognitive patterns of all students with MD, e.g., in counting, subitizing and magnitude comparison do not differ qualitatively (Kuhn et al., 2013). In our research, we address students with MD following Moser Opitz et al. (2016) and Scherer et al. (2016) as those students who encounter difficulties with a certain set of mathematical problems both on a conceptual and procedural level, including, e.g., basic arithmetic such as counting (also counting principles and counting by groups), (de-)grouping, the base-10 system, understanding place values, and basic arithmetic operations.

QUANTITY RECOGNITION

To determine quantities—i.e., to grasp a set of items and say how many they are—is a crucial skill for children to learn. Whereas young children typically already have the

ability to grasp numbers of small sets of items in one glance ("subitizing", Clements, 1999), they later on learn to count and to subitize conceptually, i.e., to make use of patterning abilities and to structure sets into subsets when determining numbers (ibid.). For students to apprehend the number range up to 20 or 100, teachers commonly use external representations such as the abacus (frame) or dot field (Gaidoschik, 2015, see Fig. 2): These representations both visualize substructures (10s, 5s, 50s)—in slightly different ways (e.g., through change of colors or gaps)—for the students to understand the base-10 system and to develop mental representations of the structures (Wartha & Schulz, 2012). Whereas investigating students' strategies in such representations (identifying what structures they use and how) is a challenging task (Obersteiner et al., 2014), researchers have found that ET may be useful to analyze students' quantity recognition strategies in structured whole number representations (Lindmeier & Heinze, 2016; Rottmann & Schipper, 2002; Schindler & Lilienthal, 2018). Lindmeier and Heinze (2016) concluded that ET data are useful to infer student strategies, and Obersteiner et al. (2014) point out that a combination of tasks on computerized versions of structured whole number representations together with ET appears to be a promising approach to assess students' strategies. For investigating students' quantity recognition strategies through ET, researchers particularly analyzed qualitatively students' scanpaths (e.g., Lindmeier & Heinze, 2016) or gaze-overlaid videos (i.e., augmented videos of the scene with the gaze visualized as point, e.g., Schindler & Lilienthal, 2018), which reveal where the students looked at and indicate student strategies. However, as Schindler and Lilienthal (2019) point out, the qualitative analysis of such gaze patterns is demanding and time-consuming. Therefore, this study uses AI, in particular pattern recognition to (partially) automate the analysis of gaze patterns and focuses on the spatial distribution of gazes over the task sheet.

THIS STUDY

Students. For answering the research question, we use data from a research project with 164 (92 males, 72 females) fifth-grade students in a German comprehensive school ("Gesamtschule"). The mean age was 10;9 (SD = 0;7) with ages ranging between 9;10 and 12;6. The participating school was in a town of 80,000 inhabitants, situated on the edge of a German urban area. The study took place in the first weeks of fifth grade.

We conducted a standardized arithmetic paper-pencil speed test (HRT; Haffner et al., 2005) with all 164 students in classroom settings. Only the first part of HRT, which can be used solely for diagnosing MD (at percentile rank (PR)<11; Haffner et al., 2005), was administered (similar to Schleifer & Landerl, 2011). The six subtests address mental addition, subtraction, multiplication, division, magnitude comparison (e.g., 7 _ 6; correct response: >) and completion tasks (e.g., _ - 2 = 6). We identified MD (at PR<11) and TD (PR>25) following the test's instructions, resulting in 69 MD students and 59 TD students. Percentile ranks between 11 and 25 are considered "at risk zone" (AR, Haffner et al., 2005, p. 20), which applied to 36 students. For the analysis of group differences, we focus on the groups of MD and TD students,

disregarding the AR students (see Fig. 1). Their mean t-values on the test were 32.1 (SD = 6.3) for the MD group, and 49.7 (SD = 7.5) for the TD group.

Tasks. We used a computerized version of the 100-bead abacus and the 100-dot field. The numbers were systematically chosen so that all ones and tens were included once. We also included 100, which led to eleven tasks (arranged according to size: 7, 15, 20, 31, 43, 54, 68, 76, 89, 92, and 100). The tasks were presented in randomized order (different randomization for each representation, i.e., abacus or dot field).

Procedure and eye-tracker. The students were tested individually in a quiet room. They were seated in front of a 24'' full HD computer monitor. We used the remote eye-tracker Tobii x3-120, which allows for video-based binocular tracking at a sampling rate of 120 Hz. Looking like a black stick, it was attached to the bottom frame of the monitor and hardly noticeable. Its presence and function was explained to the students, yet, it did not interfere with the students' work on the tasks. For adjusting the eye-tracker, a nine-point calibration was conducted. Then, before the students started working on the tasks, they first saw a picture of the respective representation (100-bead abacus, 100-dot field) and were asked to describe it. This was followed by two practice tasks with numbers that were not used in further test tasks. The students were instructed to correctly name the number of dots in every task as fast as possible. In between the tasks, the students were instructed to fixate a star in the middle of the screen before the next task appeared. The students received no response whether their answers were correct. Verbal answers were recorded through an audio-recorder.

Heat maps and spatial information. ET provides rich information and a large amount of data. The obtained gaze patterns can differ in many ways, including, in our case, where, when, for how long and in which order the students looked at the quantity recognition tasks. In order to identify group differences, we needed to choose an intermediate representation of the recorded gazes to allow for a feasible subsequent analysis. This intermediate representation should lower the dimensionality of the problem (loosely speaking, it should reduce the amount of data to be handled by the pattern recognition system) while preserving the relevant features of the gaze patterns. Following previous research that indicated that students' gaze distributions on the task sheets might differ on group level (Schindler et al., 2019), we decided for heat maps that show how gazes were spatially distributed over the presented digital task sheets. We thus disregard information about the order in which the students looked at the task and consider how long the students paid attention to certain areas only relative to the total duration of the task. To compute the individual students' heat maps, we use the Tobii Pro Lab Software and aggregate all gazes (not only fixations). We only include heat maps of correctly or inversely (common mistake in German) solved tasks for further analyses (e.g., for 68: "sixty eight" or "eighty six"), since we intend to sort out instances where students guessed rather than perceived the given information.

AI and pattern recognition methods. In order to assess which tasks' heat maps allow separating TD students from MD students, a Multivariate Analysis of Variances

(MANOVA) (Morrison, 2005) is performed. MANOVA is closely related to Linear Discriminant Analysis (LDA) (Izenman, 2013). Both methods are based on the same mathematical transformation but offer different interpretations of the results.

Unfortunately, it is not possible to use the heat maps directly for the analysis as the dimensionality of the input data that can be processed applying this method is limited by the available number of samples. Each heat map image represents a single point in a 2764800 (=1280*720*3, width*height*color) dimensional space. Therefore, it is necessary to compress the data contained in the heat maps into a lower dimensional space (fewer dimensions than students) before applying the MANOVA.

To this end, grayscale images are used, reducing dimensionality by one third. Then a Principal Component Analysis (PCA) (Abdi & Williams, 2010) is performed. PCA generates a new orthogonal coordinate system along the directions of high variance in the original data and achieves compression by dropping dimensions with the lowest variances. Intuitively speaking, those parts of the heat maps that look the same for every student, i.e., show low variance, are removed as they contain no information about group discrimination. 50 dimensions are chosen as target dimensionality in order to have at least 50 examples for both classes available, since the number of samples per class has to be higher than the dimensionality for some statistical tests in the MANOVA space. Through the use of 50 dimensions, 91% of the information was preserved on average in the compressed representation.

The heat maps, including those from AR students in the HRT test, are used to calculate the compression as the aim is to preserve all possibly occurring heat maps in the reduced space. Before the next step, the actual LDA/MANOVA, the heat maps of the AR students are removed from the data set to investigate differences between MD und TD students. The LDA/MANOVA can be understood as another compression method with the goal to maximize linear separability between the classes, i.e., TD and MD students, in the reduced space. Since we are considering a two-class problem, this step reduces to a single dimension (see the processing pipeline in Fig. 1).

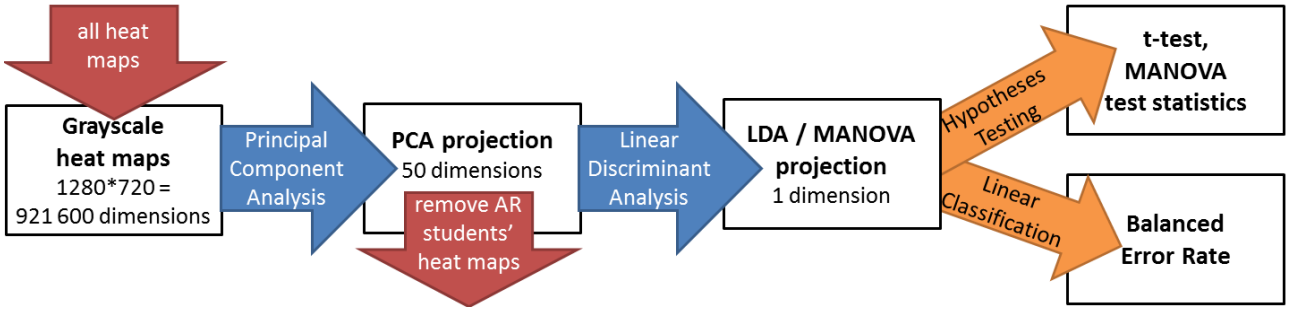


Figure 1: Pattern recognition system.

RESULTS

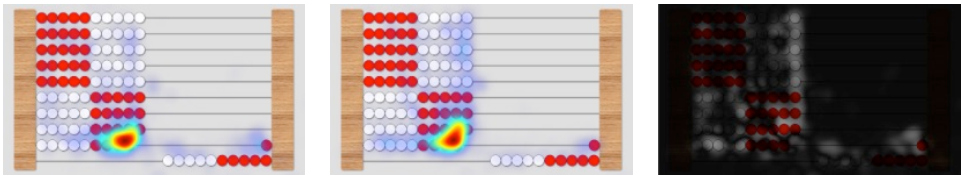
For pursuing the question if the spatial gaze patterns in quantity recognition differ between groups on task level, we used t-tests on the remaining dimension after PCA and LDA (Fig. 1). The group differences (MD vs. TD) were significant on a $p<.01$

level for all tasks, confirming significant differences between the groups' spatial gaze pattern data for each task. Yet, significant group differences do not guarantee for linear separability, since the actual performance of a classifier depends on the overlap of both data distributions. It is common practice for pattern recognition systems to report error rates as a practical assessment of how well groups can be distinguished, not just whether the differences are significant from a statistical point of view. Hence, we performed an actual classification on the given data to assess with what error rate it is actually possible to discriminate between the two groups. We report the Balanced Error Rate (BER) for a linear classifier using 3-fold cross validation (Kohavi, 1995). The BER is calculated according to $BER = \frac{1}{2}\left(\frac{fn}{fn+tp} + \frac{fp}{fp+tn}\right)$, with *fn* – number of false negatives, *tp* – number of true positives, *fp* – number of false positives, and *tn* – number of true negatives. A BER of 5% means that the average number of students wrongly classified is 5% of all samples. Any classifier that would assign classes by chance achieves a BER of 50%. The BERs are provided in Table 1.

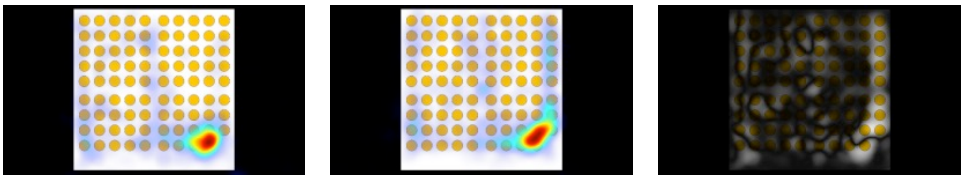
task	**7**	**15**	**20**	**31**	**43**	**54**	**68**	**76**	**89**	**92**	**100**
dot field	15.93%	18.20%	28.26%	10.19%	8.96%	16.71%	12.99%	9.42%	5.96%	18.95%	18.93%
abacus	18.05%	17.40%	14.14%	16.41%	17.32%	21.32%	4.88%	12.11%	9.73%	13.79%	17.91%

Table 1: Balanced Error Rates per task separating TD and MD students.

Certain quantities have low BERs in both representations: For these tasks, the percentage of wrongly classified students is low and the students' gazes differ substantially on group level. For 43, 68, 76, and 89 we found BERs of below 10% in at least one representation. For other quantities (e.g., 20), BERs are higher, indicating less pronounced, but still relevant differences between the groups of students in these tasks.



Abacus 89 Heat Map TD | Abacus 89 Heat Map MD | Abacus 89 Difference Map



Field 89 Heat Map TD | Field 89 Heat Map MD | Field 89 Difference Map

Figure 2: Average heat maps and difference maps for 89

Through average heat maps, the group differences in students' spatial gaze patterns can be visualized (Fig. 2; Note: Heat maps have warmer colors where the students looked at more often, and difference maps visualize the group differences of the gazes: the brighter the bigger the differences in the respective area; maps self-produced by authors). As can be seen in the examples, MD students' gazes appear to be more on the right edge of the dots/beads.

DISCUSSION

The aim of this paper was to investigate group differences in students' gaze patterns when determining quantities in structured whole number representations. We used pattern recognition from AI to find differences in students' spatial gaze patterns on the 100-abacus and dot field—in eleven tasks per representation.

Looking at statistical comparisons, we found that the spatial gaze patterns in quantity recognition on the abacus and dot field differed significantly between MD and TD students for all tasks. In every task, when determining quantities of 89, 54, or 7, the groups' gaze distributions on the (digital) task sheets were significantly different. Calculating furthermore error rates (BER in particular), we found that every task contains exploitable information to separate MD from TD students (i.e., for none of the tasks the classifier came close to 50%, i.e., guessing). In summary, through AI we found that the groups' gazes differed—substantially in some tasks. This result may hint at different strategy uses of the groups of students. This is in line with results from previous explorative, qualitative studies which revealed that MD students tend to use other strategies on such representations than TD students (Rottmann & Schipper, 2002; Schindler et al., 2019). In our study, the visualizations of average heat maps (cumulative heat map of all MD/TD students, see Fig. 2) helped to understand group differences. These visualizations shed light on the students' spatial gaze distributions on the task sheets and indicate that MD might count rows more often than TD students.

Besides these empirical findings, our results indicate what tasks (quantities) might be most adequate for identifying students with MD. Our results do not suggest that the abacus is better suited than the dot field or vice-versa, but instead certain tasks produce the lowest errors of our pattern recognition system—often in both representations. While the lower error rates (e.g., for 89) are promising, the results indicate that a highly reliable classification based on a single task is hard to achieve—and also not reasonable from a pedagogical perspective. Future research should investigate what a reasonable set of tasks may look like to perform a classification with high confidence. This would help identifying students' needs in order to support them adequately.

References

Abdi, H., & Williams, L.J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics, 2*(4), 433–459.

Gaidoschik, M. (2015). Einige Fragen zur Didaktik der Erarbeitung des „Hunderterraums". *Journal für Mathematik-Didaktik, 36*(1), 163–190.

Haffner, J., Baro, K., Parzer, P., & Resch, F. (2005). *Heidelberger Rechentest (HRT 1-4)*. Göttingen, Germany: Hogrefe.

Izenman A.J. (2013) *Linear Discriminant Analysis. Modern Multivariate Statistical Techniques.* New York: Springer.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Int. Joint Conference on Artificial Intelligence, 14*(2), 1137–1145.

Kuhn, J.T., Raddatz, J., Holling, H., & Dobel, C. (2013). Dyskalkulie vs. Rechenschwäche: Basisnumerische Verarbeitung in der Grundschule. *Lernen/Lernstörungen, 2*(4), 229–247.

Lindmeier, A., & Heinze, A. (2016). *Strategies for recognizing quantities in structured whole number representations – A comparative eye-tracking study.* Paper presented at 13th International Congress on Mathematical Education (ICME-13), 2016.

Morrison, D.F. (2005). Multivariate analysis of variance. In P. Armitage, & T. Colton (Eds.), *Encyclopedia of biostatistics.* Available at https://doi.org/10.1002/0470011815.b2a13045

Moser Opitz, E., Freesemann, O., Prediger, S., Grob, U., Matull, I., & Hußmann, S. (2017). Remediation for students with mathematics difficulties: An intervention study in middle schools. *Journal of Learning Disabilities*, *50*(6), 724-736.

Obersteiner, A., Reiss, K., Ufer, S., Luwel, K., & Verschaffel, L. (2014). Do first graders make efficient use of external number representations? The case of the twenty-frame. *Cognition and Instruction, 32*(4), 353–373.

Rottmann, T., & Schipper, W. (2002). Das Hunderter-Feld—Hilfe oder Hindernis beim Rechnen im Zahlenraum bis 100? *Journal für Mathematik-Didaktik*, *23*(1), 51–74.

Scherer, P., Beswick, K., DeBlois, L., Healy, L., & Moser Opitz, E. (2016). Assistance of students with mathematical learning difficulties: how can research support practice? *ZDM*, *48*(5), 633–649.

Schindler, M., Bader, E., Lilienthal, A.J., Schindler, F., & Schabmann, A. (2019). Quantity recognition in structured whole number representations of students with mathematical difficulties: An eye-tracking study. *Learning Disabilities: A Contemporary Journal.*

Schindler, M., & Lilienthal, A.J. (2019). Domain-specific interpretation of eye tracking data: Towards a refined use of the eye-mind hypothesis for the field of geometry. *Educational Studies in Mathematics*, 1–16 (Online First).

Schindler, M., & Lilienthal, A.J. (2018). Eye-tracking for studying mathematical difficulties, also in inclusive settings. In Bergqvist, E., Österholm, M., Granberg, C., & Sumpter, L. (Eds.), *Proc. of the 42nd Conf. of PME* (Vol.4, pp.115–122). Umeå, Sweden: PME.

Schleifer, P., & Landerl, K. (2011). Subitizing and counting in typical and atypical development. *Developmental Science*, *14*(2), 280–291.

Wartha, S., & Schulz, A. (2012). *Rechenproblemen vorbeugen. Grundvorstellungen aufbauen. Zahlen und Rechnen bis 100.* Berlin: Cornelsen.

World Health Organisation (WHO) (Ed.) (2018). *ICD-11: International classification for diseases for mortality and morbidity statistics (ICD-11 MMS) 2018 version.* http://www.who.int/classifications/icd/. Accessed 2018-09-10.