

Intraindividual Variability in the Spotlight of Special Education Research

Carmen Zurbriggen & Julia Winkes (Eds.)



Empirische Sonderpädagogik

17. Jahrgang · Heft 3 · 2025

Intraindividual Variability in the Spotlight of Special Education Research

Carmen Zurbriggen & Julia Winkes (Eds.)

Inhalt

- 179 Intraindividual Variability in the Spotlight of Special Education Research – Editorial to the Special Issue
Carmen Zurbriggen & Julia Winkes
- 184 The Power of Randomization Tests to Detect Individual Intervention Effects: Insights from a Monte Carlo Study
Jürgen Wilbert, Moritz Börnert-Ringleb & Timo Lüke
- 200 Heterogenität im Individuum: die Bedeutung von intraindividueller Variabilität aus der Perspektive der Sonderpädagogik
Julia Winkes
- 215 Intra-individual Differences in Situational Appraisals of Mathematics Anxiety in Grade Four Students
Moritz Herzog, Miriam Balt & Carmen Zurbriggen
- 232 Welche Rolle spielt das Fördersetting? Längsschnittstudie zur Wirkung der Förderung in der Grafomotorik
Eckhart Michael, Schwery Klingele Joséphine, Sägesser Wyss Judith & Wüthrich Sergej

Impressum

Herausgeber:

PROF. DR. MATTHIAS GRÜNKE
Universität zu Köln
Humanwissenschaftliche Fakultät
Department Heilpädagogik und
Rehabilitation
Klosterstr. 79b
50931 Köln
matthias.gruenke@uni-koeln.de

PROF. DR. TIMO LÜKE
Karl-Franzens-Universität Graz
Inklusive Bildung und Unterrichtsentwicklung
Merangasse 70/II
8010 Graz, Österreich
timo.lueke@uni-graz.at

PROF. DR. JÜRGEN WILBERT
Universität Münster
Professur für Erziehungswissenschaft mit dem
Schwerpunkt Sonderpädagogik im Kontext
inklusive Bildung
Georgskommende 26
48143 Münster
juergen.wilbert@uni-muenster.de

Fachbeirat:

Dr. Stefan Blumenthal, Universität Rostock
Prof. Dr. Moritz Börnert-Ringleb,
Universität Hannover
Prof. Dr. Gino Casale, Universität Wuppertal
Prof. Dr. Armin Castello, Universität Flensburg
Prof. Dr. Jürgen Cholewa, Pädagogische
Hochschule Heidelberg
Prof. Dr. Michael Grosche,
Universität Wuppertal
Prof. Dr. Thomas Hennemann,
Universität zu Köln
Prof. em. Dr. Manfred Hintermair,
Pädagogische Hochschule Heidelberg
Prof. Dr. Christian Huber, Universität Wuppertal
Dr. Johanna Krull, Universität zu Köln
Prof. Dr. Jan Kuhl,
Technische Universität Dortmund
Prof. Dr. Christian Liesen,
Hochschule für Heilpädagogik Zürich
Prof. Dr. Friedrich Linderkamp,
Universität Wuppertal
Prof. Dr. Reto Luder,
Pädagogische Hochschule Zürich
Prof. Dr. Carina Lüke, Universität Würzburg
Prof. Dr. Claudia Mähler, Universität Hildesheim
Prof. Dr. Elisabeth Moser Opitz,
Universität Zürich
Prof. Dr. Christoph Michael Müller,
Universität Fribourg
Prof. em. Dr. Susanne Nußbeck,
Universität zu Köln

Prof. Dr. Christoph Ratz, Universität Würzburg
Prof. Dr. Gabriele Ricken, Universität Hamburg
Prof. Dr. Christian Rietz,
Pädagogische Hochschule Heidelberg
Prof. Dr. Alfred Schabmann, Universität zu Köln
Prof. Dr. Roland Stein, Universität Würzburg
Prof. Dr. Mark Stemmler,
Universität Erlangen-Nürnberg
Prof. Dr. Karolina Urton, Universität Münster
Prof. Dr. Jürgen Walter, Universität Flensburg
Prof. em. Dr. Franz B. Wember,
Technische Universität Dortmund
Prof. Dr. Carmen Zurbriggen,
Universität Fribourg

Verlag:

Pabst Science Publishers
Eichengrund 28
D-49525 Lengerich/Westfalen
Tel. +49(0) 5484-308
Fax +49(0) 5484-550
E-Mail: pabst@pabst-publishers.com
www.pabst-publishers.com
www.psychologie-aktuell.com

Erscheinungsweise:

4xjährlich

Preise:

(Jahresabonnement: 40,- €
Einzelheft: 15,- €
inkl. MwSt. und Versand)

Verlagsredaktion:

Erika Wiedenmann
wiedenmann@pabst-publishers.com

Herstellung:

Patrick Orths
orths@pabst-publishers.com

Administration:

Karin Moenninghoff
moenninghoff@pabst-publishers.com

Druck:

KM-Druck 2.0, D-64823 Groß-Umstadt

ISSN 1869-4845 (Printausgabe)

ISSN 1869-4934 (ebook)

Intraindividual Variability in the Spotlight of Special Education Research – Editorial to the Special Issue

Carmen Zurbriggen & Julia Winkes

Department of Special Education, University of Fribourg, Switzerland

Fluctuations in behaviour or short-term variations in a person's emotional or cognitive state have long been of interest in empirical research (Eid & Diener, 1999; Fiske & Rice, 1955). Beyond attempts to statistically control these fluctuations or identify stable effects, the focused study of intraindividual variability across different phenomena has gained prominence in recent years (Oeltjen et al., 2023; Schmiedek et al., 2009).

Previous research has emphasised the importance of considering intraindividual variability in cognitive functioning among both children with developmental disorders (Borella et al., 2011) and older adults (Ghisletta et al., 2013). One study demonstrated that substantial day-to-day and moment-to-moment fluctuations in working memory were associated with mathematics and reading performance in schoolchildren (Dirk & Schmiedek, 2016). The fluctuating nature of emotions and its pivotal role in mental health and well-being has also been evidenced in numerous studies (Houben et al., 2015). The momentum in intraindividual variability research has been driven by significant advances in data collection methods and analytical approaches (Hamaker & Wichers, 2017; Oeltjen et al., 2023), enabling novel insights.

However, in the field of special education, intraindividual variability has received little attention, although fluctuations in individual competencies and highly variable emotional and social experiences or behaviours are common. Nevertheless, methods capable of capturing intraindividual variability, such as experience sampling method or intensive longitudinal methods, are increasingly being applied in special education research (Venetz & Zurbriggen, 2015; Zurbriggen et al., 2018). More application-oriented approaches, including curriculum-based measurement, direct behaviour rating, or (controlled) single-case studies, also generate suitable data to investigate intraindividual variability. Yet, during analysis, such within-person differences and fluctuating processes are rarely modelled.

With this in mind, in January 2024 we, as guest editors of the journal *Empirische Sonderpädagogik*, invited submissions focusing on intraindividual variability in special education. Although the call for papers generated considerable interest, as evidenced by numerous verbal and written responses from colleagues, there was no surge of abstract submissions for relevant articles. This can be interpreted, at least in part, as an indication of the aforementioned research gap. Nevertheless, the three papers that successfully passed both the abstract selection and peer review processes, and were ultimately included in the special issue, clearly demonstrate the particular relevance of considering intraindividual variability in empirical special education research.

The first paper, by Jürgen Wilbert, Moritz Börner-Ringleb, and Timo Lüke, begins with the observation that, although there is general consensus on the usefulness of single-case

experimental designs in special education research, there remains no agreement on the most appropriate analytical approach. Their contribution, *Randomization tests for detecting intraindividual variability in single cases: An analysis of power and alpha-error probability*, examines how different design conditions affect the performance of randomisation tests in detecting meaningful within-person changes. Their findings show that, under certain conditions, randomisation tests achieve high statistical power and low alpha error rates. The paper places particular emphasis on their robustness to trends and their applicability in multiple-case designs, even when the number of measurements is relatively small. As such, this methodological paper provides valuable insights for further studies employing single-case experimental designs.

The second paper, by Julia Winkes, entitled *Heterogenität im Individuum: die Bedeutung von intraindividueller Variabilität aus der Perspektive der Sonderpädagogik* (*Heterogeneity in the Individual: The Significance of Intra-Individual Variability from the Perspective of Special Education*), highlights that intra-individual variability—or fluctuations in a person's behaviour and performance over time—should not be dismissed as mere measurement error, but instead recognised as a valuable source of information. The article calls for a shift in perspective towards individual developmental processes and dynamic analyses, supported by theoretical approaches such as complex dynamic systems theory and methodological approaches such as single-case studies or the experience sampling method. This perspective provides new impetus not only for empirical research in special education but also for diagnosis and support, particularly with respect to the adaptive instruction of learning processes.

The third paper, by Moritz Herzog, Miriam Balt, and Carmen Zurbriggen, presents an original empirical study entitled *Intra-individual differences in situational appraisals of mathematics anxiety in grade four students*. Employing the experience sampling method, the study explores how subjectively perceived teaching characteristics affect the experience of mathematics anxiety in fourth graders, with a special focus on gender-specific differences. The results reveal that higher task difficulty is associated with increased mathematics anxiety, while the perceived relevance of the situation has a gender-specific effect. The study underlines the importance of context-sensitive survey methods and intraindividual analyses for understanding mathematics anxiety in the classroom.

Taken together, the three papers in this special issue on intraindividual variability provide significant contributions and points of reference for future studies in empirical special education research—aptly reflecting the motto: “Small but meaningful”.

Intraindividuelle Variabilität im Fokus sonderpädagogischer Forschung – Editorial zum Themenheft

Schwankungen im Verhalten oder kurzfristig variierende emotionale und kognitive Zustände einer Person sind in der empirischen Forschung bereits seit längerem von Interesse (Eid & Diener, 1999; Fiske & Rice, 1955). Neben Bestrebungen, solche Fluktuationen statistisch zu kontrollieren oder nach stabilen Effekten zu suchen, ist die gezielte Untersuchung intraindividueller Variabilität von verschiedenen Phänomenbereichen in den letzten Jahren zunehmend in den Vordergrund gerückt (Oeltjen et al., 2023; Schmiedek et al., 2009).

So verweisen bisherige Studien etwa auf die Bedeutung der intraindividuellen Variabilität im Bereich kognitiver Funktionen bei Kindern mit Entwicklungsstörungen (Borella et

al., 2011) oder im höheren Erwachsenenalter (Ghisletta et al., 2013). In einer Studie aus dem schulischen Kontext zeigten sich, dass substanzielle ‘Tag-zu-Tag’ und ‘Moment-zu-Moment’ Fluktuationen im Arbeitsgedächtnis mit Mathematik- und Leseleistungen verknüpft waren (Dirk & Schmiedek, 2016). Die entscheidende Bedeutung des fluktuierenden Charakters von Emotionen für die psychische Gesundheit und das Wohlbefinden wurde in etlichen Studien belegt (Houben et al., 2015). Besonderen Aufschwung erhalten hat die Forschung rund um intraindividuelle Variabilität durch substanzielle Fortschritte von Datenerhebungsverfahren und Analyseansätzen (Hamaker & Wichers, 2017; Oeltjen et al., 2023), die dadurch neuartige Erkenntnisse ermöglichen.

In der empirischen sonderpädagogischen Forschung stand die intraindividuelle Variabilität bislang eher selten im Fokus des Interesses, obgleich Schwankungen von individuellen Kompetenzen und teilweise stark variierende emotionale und soziale Erlebens- oder Verhaltensweisen auch hier allgegenwärtige Phänomene darstellen. Zunehmend kommen in der sonderpädagogischen Forschung jedoch Erhebungsmethoden wie die Experience Sampling Method bzw. Intensive Longitudinal Methods zum Einsatz, welche die intraindividuelle Variabilität zu erfassen vermögen (Venetz & Zurbriggen, 2015; Zurbriggen et al., 2018). Auch die stärker anwendungsbezogene Forschung, die Methoden verwendet wie etwa Lernverlaufsdiagnostik, Direct Behavior Rating oder (kontrollierte) Einzelfallanalyse, generiert Daten, die eine Fokussierung auf die intraindividuelle Variabilität nahelegen. Bei den Analysen werden diese in der Person liegenden Unterschiede und fluktuierenden Prozesse allerdings noch vergleichsweise wenig entsprechend modelliert.

Vor diesem Hintergrund haben wir als Gastherausgeberinnen im Januar 2024 dazu eingeladen, Beiträge mit einem Fokus auf intraindividuelle Variabilität und von sonderpädagogischer Relevanz bei der Zeitschrift *Empirische Sonderpädagogik* einzureichen. Während der Call bzw. Aufruf für den Themenschwerpunkt zwar auf reges Interesse stieß – was sich in etlichen mündlichen oder schriftlichen Rückmeldungen zeigte – blieb ein Ansturm an Einreichungen von Abstracts für entsprechende Beiträge hingegen aus. Dies kann zumindest teilweise als Indiz für die bereits erwähnte Forschungslücke interpretiert werden. Nichtsdestotrotz zeigen die drei Beiträge, die sowohl dem Auswahlverfahren der Abstracts als auch dem Peer-Review-Prozess standhielten und schlussendlich in den Themenschwerpunkt aufgenommen werden konnten, (bereits) recht deutlich die besondere Relevanz der Berücksichtigung von intraindividueller Variabilität in der empirisch sonderpädagogischen Forschung auf.

Ausgangspunkt des Beitrags von Jürgen Wilbert, Moritz Börnert-Ringleb und Timo Lücke ist, dass trotz breiter Zustimmung zur Relevanz und Nützlichkeit von Einzelfallexperimenten (single-case experimental designs) für die sonderpädagogische Forschung bislang kein Konsens über den am besten geeigneten Analyseansatz besteht. Der Beitrag mit dem Titel *Randomization tests for detecting intraindividual variability in single-cases: An analysis of power and alpha-error probability* wird anhand von Monte-Carlo-Simulationen untersucht, wie sich verschiedene Designbedingungen auf die Performanz von Randomisierungstests bei der Ermittlung von bedeutsamen Veränderungen innerhalb einer Person (within-person changes) auswirken. Die Ergebnisse zeigen, dass Randomisierungstests unter bestimmten Bedingungen eine hohe statistische Power und geringe Alpha-Fehlerraten aufweisen. Besonders hervorgehoben wird die Robustheit gegenüber Trends und das Potenzial in Mehrfachfalldesigns (multiple-case designs), selbst bei Designs mit relativ wenigen Messungen. Dieser Forschungsmethodenbeitrag liefert damit wichtige Anhaltspunkte für die weitere Arbeit mit single-case experimental designs.

Der Beitrag von Julia Winkes mit dem Titel *Heterogenität im Individuum: Die Bedeutung von intraindividueller Variabilität aus der Perspektive der Sonderpädagogik* verdeutlicht,

dass intraindividuelle Variabilität – respektive Schwankungen im Verhalten und in der Leistung einer Person über die Zeit – nicht (nur) als Messfehler, sondern als wertvolle Informationsquelle betrachtet werden sollte. Der Beitrag plädiert für einen Perspektivenwechsel hin zu individuellen Entwicklungsprozessen und dynamischen Analysen, gestützt durch theoretische Ansätze wie die Complex Dynamic Systems Theorie und methodische Zugänge wie Single-Case Studies oder die Experience Sampling Method. Daraus ergeben sich neue Impulse nicht nur für die empirische sonderpädagogische Forschung, sondern auch für die Diagnostik und Förderung, insbesondere im Hinblick auf die adaptive Gestaltung von Lernprozessen.

Der Beitrag von Moritz Herzog, Miriam Balt und Carmen Zurbriggen entspricht einer empirischen Originalarbeit und trägt den Titel *Intra-individual differences in situational appraisals of mathematics anxiety in grade four students*. Die darin berichtete Studie untersucht mithilfe der Experience Sampling Method Effekte von subjektiv wahrgenommenen Unterrichtsmerkmalen auf das Erleben von Mathematikangst bei Viertklässler:innen unter besonderer Berücksichtigung geschlechtsspezifischer Unterschiede. Die Ergebnisse zeigen, dass eine höhere Aufgabenschwierigkeit mit stärkerer Mathematikangst verknüpft ist, während die wahrgenommene Relevanz der Situation geschlechtsspezifisch wirkt. Die Ergebnisse unterstreichen die Bedeutung kontextsensitiver Erhebungsmethoden und intraindividuell Analysen für das Verständnis von Mathematikangst im Unterricht.

Gesamthaft betrachtet liefern die drei Beiträge des Themenschwerpunktes zu *intraindividuell Variabilität* wertvolle Anregungen oder Anhaltspunkte für weitere Studien im Bereich der empirischen sonderpädagogischen Forschung – ganz im Sinne von: «Klein aber fein».

References

- Borella, E., Chicherio, C., Re, A. M., Sensini, V. & Cornoldi, C. (2011). Increased intraindividual variability is a marker of ADHD but also of dyslexia: A study on handwriting. *Brain and Cognition*, 77(1), 33–39. <https://doi.org/10.1016/j.bandc.2011.06.005>
- Dirk, J., & Schmiedek, F. (2016). Fluctuations in elementary school children's working memory performance in the school context. *Journal of Educational Psychology*, 108(5), 722–739. <https://doi.org/10.1037/edu0000076>
- Eid, M., & Diener, E. (1999). Intraindividual variability in affect: Reliability, validity, and personality correlates. *Journal of Personality and Social Psychology*, 76(4), 662–676. <https://doi.org/10.1037/0022-3514.76.4.662>
- Fiske, D. W., & Rice, L. (1955). Intra-individual response variability. *Psychological Bulletin*, 52(3), 217–250. <https://doi.org/10.1037/h0045276>
- Ghisletta, P., Fagot, D., Lecerf, T., & de Ribaupierre, A. (2013). Amplitude of fluctuations and temporal dependency in intraindividual variability: Prediction of cognitive performance two years later. *GeroPsych: The Journal of Gerontopsychology and Geriatric Psychiatry*, 26(3), 141–151. <https://doi.org/10.1024/1662-9647/a000091>
- Hamaker, E. L., & Wichers, M. (2017). No time like the present: discovering the hidden dynamics in intensive longitudinal data. *Current Directions in Psychological Science*, 26(1), 10–15. <https://doi.org/10.1177/0963721416666518>
- Houben, M., Van Den Noortgate, W., & Kuppens, P. (2015). The relation between short-term emotion dynamics and psychological well-being: A meta-analysis. *Psychological Bulletin*, 141(4), 901–930. <https://doi.org/10.1037/a0038822>
- Oeltjen, L., Koch, T., Holtmann, J., Münch, F. F., Eid, M., & Nussbeck, F. W. (2023). A general framework for the inclusion of time-varying and time-invariant covariates in latent state-trait models. *Psychological Methods*, 28(5), 1005–1028. <https://doi.org/10.1037/met0000592>
- Schmiedek, F., Lövdén, M., & Lindenberger, U. (2009). On the relation of mean reaction time and intraindividual reaction time variability. *Psychology and Aging*, 24(4), 841–857. <https://doi.org/10.1037/a0017799>
- Venetz, M. & Zurbriggen, C. (2015). Intensive

longitudinal methods – ihre Eignung für die sonderpädagogische Forschung und exemplarische Anwendungsmöglichkeiten. *Empirische Sonderpädagogik*, 7(3), 194–205. <https://doi.org/10.25656/01:11381>

Zurbriggen, C. L. A., Venetz, M., & Hinni, C. (2018). The quality of experience of students with and without special educational needs in everyday life and when relating to peers. *European Journal of Special Needs Education*, 33(2), 205–220. <https://doi.org/10.1080/08856257.2018.1424777>

Author Note

 Carmen Zurbriggen
<https://orcid.org/0000-0002-0906-2000>

 Julia Winkes
<https://orcid.org/0000-0002-4383-1926>

Corresponding author

Carmen L. A. Zurbriggen
University of Fribourg
21 Rue St. Pierre-Canisius
CH-1700 Fribourg
carmen.zurbriggen@unifr.ch

The Power of Randomization Tests to Detect Individual Intervention Effects: Insights from a Monte Carlo Study

Jürgen Wilbert¹, Moritz Börnert-Ringleb² & Timo Lüke³

¹ University of Münster

² University of Hannover

³ University of Kassel

Abstract

Despite an overall agreement on the usefulness of single-case experimental designs (SCED), there is no consensus on the most suitable approach to analyzing the resulting single-case data. Previous research indicates that reliable and rigorous analysis methods in SCED are crucial for the sensitivity, specificity, and robustness of subsequent inferences on the effectiveness of interventions. Among others, randomization tests depict a promising approach to data analysis, which received less attention in research and practice than more popular analytical approaches. So far, no clear picture of the exact conditions under which randomization tests yield high statistical power and low alpha-error rates has emerged. In the present paper, we applied several Monte Carlo studies to scrutinize the influence of various design conditions on the sensitivity and specificity of a randomization test for detecting meaningful within-person changes. We focused on the number of measurement times of each phase, the size of the intervention effect, the reliability of the measurements, the trend effect, and the number of cases in multiple-case designs. The results of our study demonstrate that a) the power of randomization tests for analyzing single-case data is high and the alpha error rate is low when cases have at least ten measurements in Phase A and the intervention effect is large, b) randomization tests are robust against data trends, and c) multiple-case designs rapidly increase the power, allowing even for minimal designs with 15 or fewer measurements.

Keywords: Single-Case Experimental Design, Intraindividual Variability, Randomization Test, Monte Carlo Simulation, Special Education

Die Teststärke von Randomisierungstests zur Identifikation intraindividueller Interventionseffekte: Erkenntnisse aus einer Monte-Carlo-Studie

Zusammenfassung

Trotz eines breiten Konsenses über die Nützlichkeit von Einzelfall-Experimentaldesigns (Single-Case Experimental Designs, SCED) besteht bislang keine Einigkeit über den am besten geeigneten Ansatz zur Analyse der dabei entstehenden Einzelfalldaten. Frühere Forschung

zeigt, dass verlässliche und methodisch stringente Analysemethoden in SCED entscheidend für die Sensitivität, Spezifität und Robustheit der Schlussfolgerungen zur Wirksamkeit von Interventionen sind. Unter den möglichen Verfahren stellen Randomisierungstests einen vielversprechenden Analyseansatz dar, der in Forschung und Praxis bislang weniger Beachtung gefunden hat. Bisher fehlte jedoch ein klares Bild der Bedingungen, unter denen Randomisierungstests eine hohe Teststärke (Power) und niedrige α -Fehler-Raten aufweisen. In der vorliegenden Arbeit wurden mehrere Monte-Carlo-Studien durchgeführt, um den Einfluss verschiedener Designbedingungen auf die Sensitivität und Spezifität eines Randomisierungstests zur Erkennung bedeutsamer intraindividuelle Veränderungen zu untersuchen. Im Fokus standen dabei die Anzahl der Messzeitpunkte je Phase, die Größe des Interventionseffekts, die Reliabilität der Messungen, Trendverläufe sowie die Fallzahl in Mehrfäledesigns. Die Ergebnisse zeigen, dass (a) die Power von Randomisierungstests zur Analyse von Einzelfalldaten hoch und die α -Fehler-Rate gering ist, wenn pro Phase mindestens zehn Messungen vorliegen und der Interventionseffekt groß ist, (b) Randomisierungstests robust gegenüber Datentrends sind und (c) Designs mit mehreren Fällen die Power deutlich erhöhen, sodass selbst Minimaldesigns mit 15 oder weniger Messungen aussagekräftig werden.

Schlagerwörter: Einzelfall-Experimentaldesign, intraindividuelle Variabilität, Randomisierungstest, Monte-Carlo-Simulation, Sonderpädagogik

Special education research often aims to gain insights into the effectiveness of educational interventions. In doing so, researchers are confronted with the challenge of the relatively small populations, limiting access to large samples. In addition, one might be particularly interested in research that examines the effectiveness of interventions under varying real-world conditions, which complicates the implementation of large-scale evaluative designs. Especially in the early and exploratory stages of research, it is reasonable to apply research designs that focus on small samples and promise to yield high internal validity when it comes to causal inference. Here, one frequently used group of designs that meets these requirements is 'Single-Case Experimental Designs' (SCED; Ledford et al., 2022). In SCED, intensive data on the development of at least one individual ($N = 1$) are collected over time across different phases, e.g., using curriculum-based measurement (CBM) or experience sampling methods (ESM). In the most basic variant of SCED, the AB design, the individual development during an intervention (phase B) is compared against the

development in a baseline (phase A). Despite an overall agreement on the usefulness of such SCED, there is no consensus on the most suitable approach to analyzing the resulting single-case data (Manolov & Moeyaert, 2017). At the same time, previous research indicates that reliable and rigorous analysis methods in SCED are crucial for the sensitivity, specificity, and robustness of subsequent inferences on the effectiveness of interventions (e.g., Wilbert et al., 2022). Among others, randomization tests depict a promising approach to data analysis, which did not receive equal attention in research and practice compared to more popular analytical approaches (e.g., visual inspection, overlap indices). So far, insights into the sensitivity, specificity, and robustness of randomization tests are lacking, especially regarding the influence of design parameters such as design length, intervention effect, and data reliability.

Single-Case Experimental Designs (SCEDs)

Single-Case Experimental Designs (SCEDs) are commonly used in educational research. They allow us to gain insights into the effectiveness of interventions in small samples of students. While the external validity of an individual study is limited, SCEDs yield high internal validity. The variations of SCED are based on a simple AB design, which is the building block of more complex extensions. These designs are unified by a) the intensive collection of individual data over time and b) the deliberate integration of an intervention during that period. In basic AB designs, data collected during an intervention phase (phase B) are compared to data collected before the intervention (phase A; Baseline). More complex design variations build up on these basic AB designs by adapting the number and order of phases. In alternating SCED, baseline and intervention phases are repeated multiple times. In Multiple Baseline Designs (MBDs), AB designs are extended/replicated across settings or individuals. Such design extensions allow for the generalization of findings across individuals/settings, increasing the validity of conclusions. The variations of SCED are commonly applied in special education research, e.g., to examine the effectiveness of reading interventions (e.g., Sperling et al., 2019), math interventions (e.g., Herzog & Casale, 2024), behavioral interventions (e.g., Briesch & Briesch, 2016) as well as classroom management techniques (e.g., Soares et al., 2016). At the same time, not all applied designs have equal quality. Even studies that use the same general approach of SCED differ with regard to the exact design characteristics, such as the number of measurements and scaling of the dependent variable and method of data analysis. This, however, is important to consider, as the statistical power of the individual study depends on both aspects (e.g., Wilbert et al., 2022).

(Statistical) Analysis of SCEDs

Different approaches exist in the analysis of data from SCED. These range from visual analysis, using descriptive effect size measures of (non)-overlap indices, to more advanced statistical methods of (multilevel) piecewise regression models and randomization tests. The first and most traditional approach to analyzing single-case data is the visual analysis of a single-case data plot (Matyas & Greenwood, 1990; Parker & Brossart, 2003). Here, the researcher compares the features of different phases with regard to the potential effectiveness of the intervention. In their judgment, researchers might focus on various characteristics of the single-case plot, such as a potential level effect, a trend effect, the immediacy of change, and (non)-overlap between the different phases. Visual elements, such as lines representing the phase mean or median, can support visual inspection. Visual analysis relies on perception and is therefore prone to potential misinterpretations. It has been described that aspects such as data variability (outliers) and data trends impact the validity of the judgment (Klapproth, 2018; Wilbert et al., 2021). Visual analysis is hence often accompanied by the use of (non)-overlap indices. Such indices depict an umbrella term for a range of measures developed in the last decades to quantify the potential effectiveness of interventions in single-case data (Parker, Vannest & Davis, 2011). Common variants comprise the family of Tau-U indices (Parker et al., 2011), the percentage exceeding the median (PEM; Ma, 2006), and the percentage of non-overlapping data (PND; Scruggs et al., 1987). Depending on the exact variant, data points in phase B are usually compared to one or all data points in phase A. Depending on the overlap of both phases, one would argue for the strength of the intervention. At the same time, overlap indices are especially prone to the effects of data trends (e.g., Wilbert, Lüke, & Börnert-Ringleb, under review). More advanced statistical techni-

ques are being applied to overcome the limitations of more basic approaches, such as the proneness to autocorrelation and data trends. These more sophisticated approaches comprise piecewise regression models (Wilbert, Börnert-Ringleb, & Lüke, 2022) and randomization tests (e.g., Edgington & Onghena, 2007). With regard to randomization tests, Heyvaert and Onghena (2014) argue that “statistical significance tests produce consistent results that are independent of the performer of the analysis, they can be applied with unstable baseline data and with increased intrasubject variability, [...]” (p. 513).

Randomization Tests

As a result, randomization tests are frequently discussed as a promising approach to analyzing SCED (Heyvaert & Onghena, 2014). The basic idea of a randomization

test is to think counterfactually: Assuming that where the phases change had no effect on the measured data, what would be the difference between the phases of my case if I had started phase B at a different time? Given the possible phase differences under the assumption that the phase change had no effect, what is the likelihood of the actual phase differences of the original case?

Therefore, a number of new cases are generated with a random start for each phase. This means these new cases have the same data as the original but different starting points for each phase. A specific outcome measure e.g., the mean difference between Phase A and Phase B data) is now calculated for each new case (see an illustration in Figure 1). When enough random cases have been generated, we calculate this outcome measure for each case (e.g., the mean difference between phase B and phase A data). The outcome for the original case is

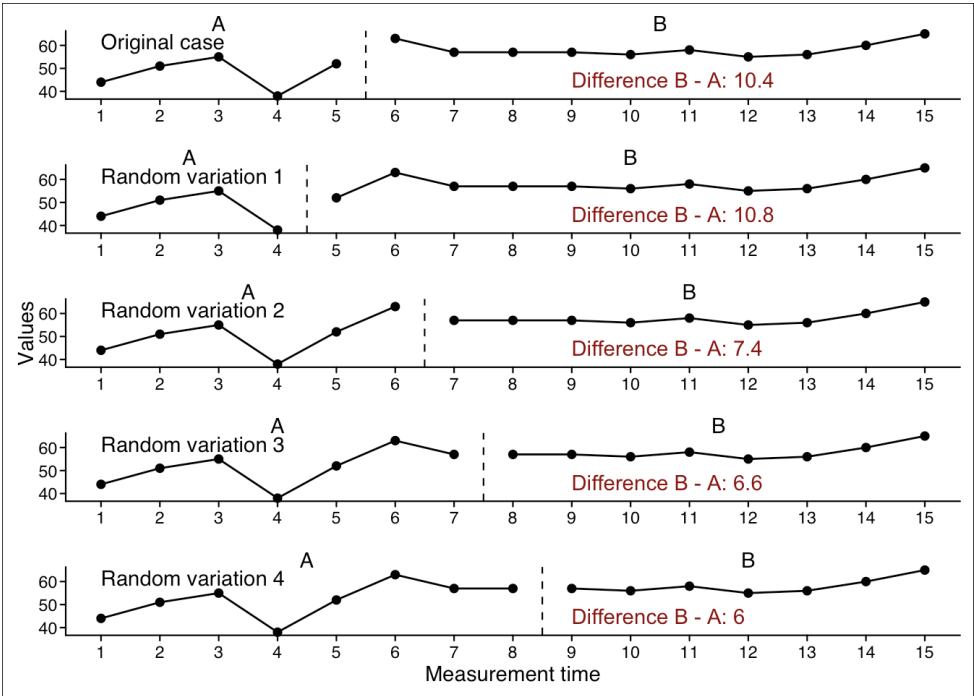


Figure 1
Illustration of the randomization of a single-case study and the resulting distribution of mean differences

now compared to the collection of outcome measures. The percentile of the original statistic within the newly generated statistic values is the probability of the original outcome assuming a random distribution of starting points for each phase. This percentile is returned as the p-value in the randomization test analyses (Heyvaert & Onghena, 2014).

In the case of several single-cases within one study - for example, in a multiple baseline design - the same procedure applies. Here, the starting points of each phase are randomly permuted for each individual case, and an aggregated statistic (e.g., the difference between all Phase B and Phase A values across all cases) is compared to the statistic of the actual phase constellation.

Statistical Power and Alpha-Error Rates of Randomization Tests

Despite the frequent use of randomization tests in educational research, limitations of the method are outlined by Heyvaert and Onghena (2014): A first limitation refers to the aspect that the beginning of the intervention has to be determined randomly before the start of data collection. This contrasts with the potential need to determine a stable baseline before introducing the method of interest ("response-guided experimentation"; Heyvaert and Onghena, 2014). Secondly, the authors highlight that the statistical power to correctly identify interventions' effects depends on a range of variables (e.g., design characteristics and effect size). Among other, the design characteristics that were focussed were design length (e.g., Ferron & Ware, 1995; Ferron & Onghena, 1996; Ferron & Sentowich, 2002; Manolov & Solanas, 2009), effect size (e.g., Ferron & Sentowich, 2002; Levin et al., 2014; Manolov et al., 2010; Manolov & Solanas, 2009), and auto-correlation (e.g., Lall & Levin, 2004; Manolov et al., 2010). Not all of these simulations focussed on the alpha-error probability (Type 1 error rates) but highlighted the statistical power. At the

same time, differences between existing simulation studies arise regarding the focused single-case design (e.g., ABAB, multiple baseline) and the randomization procedure applied, so results cannot always be compared one to one. In addition, to our knowledge, specific design characteristics have not been focussed in detail in existing simulation studies, which is particularly true for the reliability of the measurement. It becomes apparent that so far no clear picture on the exact conditions under which randomization tests yield high statistical power and alpha-error rates emerges.

Research questions (RQ)

In this paper, we investigate several of the desiderata mentioned above regarding the power and alpha-error probability when applying a randomization test for single-case studies under various design specifications. Specifically, we will investigate the impact of

1. the number of measurement times of each phase relative to the size of the intervention effect (RQ1),
2. the reliability of the measurements (i.e., the extent of measurement error) (RQ2),
3. the presence and magnitude of a baseline trend effect. (RQ3),
4. the number of cases in multiple-case designs (RQ4)

on both statistical power and alpha-error rates.

Method

We employed several Monte Carlo studies to investigate the research questions. By that means, we aim to artificially generate a large number of single-case studies under pre-defined specifications and to reanalyze these cases using a randomization test in a second step. This will allow us to identify the frequency of correctly detected existing effects and falsely detected non-existing effects. The proportion of correctly detected

effects is the power, and the proportion of falsely detected effects is the alpha error probability of the randomization test under the given specification.

Simulation designs

We set up four simulation studies to address the research question (see Table 1 for the exact specifications and iterations). In simulation 1, we independently varied the length of phase A, phase B, and the intervention effect size (level-effect d) to address the first research question. In simulation 2, we independently varied the reliability of the measurements, the overall number of measurements, and the intervention effect size (level-effect d) to answer research question 2. In simulation 3, we independently varied the size of a baseline trend effect, the overall number of measurements, and the size of the intervention effect (level-effect d) to address research question 3. Lastly, in simulation 4, we independently varied the number of cases in a multiple-case design, the overall number of measurements, and

the intervention effect size (level-effect d) to focus research question 4.

Case Construction

For each of the resulting specifications, 10,000 cases were generated. A random single-case study was created by calculating the expected values for the dependent variable for each measurement. Based on the T-score scale, we assumed the dependent variable to be on an integer scale with a mean of 50 and a standard deviation of 10. For an illustration of an exemplary randomly generated single-case, see Figure 2.

The formula for calculating a score was:

(1)
$$y_i = start + phase_i \times level + mt_i \times trend + phase_i \times (mt_i - nA) + e_i$$

where:

y := The predicted score rounded to an integer; i := The index of a measurement.

Table 1
Simulation parameters

Simulation	Number of measurements	Level-effect d	Trend-effect d	Reliability r_{tt}	Number of cases	Specifications
#1	Phase A: {5, 10, 15, 20}	{0.5, 1.0, 1.5, 2.0}	0	0.8	1	80
#2	Phase B: {10, 15, 20, 30, 40} {15, 30, 45, 60} with Phase A 1/3 and phase B 2/3 of the measurements	{0.5, 1.0, 1.5, 2.0}	0	{0.6, 0.7, 0.8, 0.9}	1	64
#3	{15, 30, 45, 60} with Phase A 1/3 and phase B 2/3 of the measurements	{0.5, 1.0, 1.5, 2.0}	{-0.50, -0.25, 0.00, +0.25, +0.5}	0.8	1	80
#4	{9, 15, 30, 45} with Phase A 1/3 and phase B 2/3 of the measurements	{0.5, 1.0, 1.5, 2.0}	0	0.8	{1, 3, 5, 7}	64

Note. The slope effect was $d = 0$ for all simulations. Vectors in curly brackets indicate iterations.

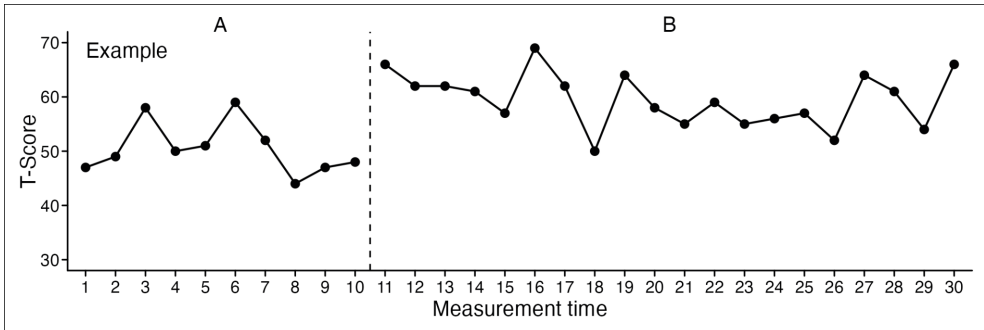


Figure 2

Example case for a specification with a level effect of $d = 1$ and a reliability of $r_{tt} = 0.8$

start := The initial score at the start of the study drawn from a normal distribution $N(M = 50, SD = 10)$.

phase := A variable with 0 for phase A and 1 for phase B measurements.

level := The change due to the intervention effect in standardized mean differences d .

mt := The measurement time (1, 2, ..., n)

trend := A trend effect leading to a continuous change for each measurement. The scale is the changes in standardized mean differences d per measurement time.

nA := The number of measurements of phase A.

e := The measurement error randomly drawn from a normal distribution with $M =$

0 and $SD = \sqrt{\frac{(1-r_{tt})}{r_{tt}} \times SD_{start}^2}$. Here, r_{tt} is the assumed reliability of the measurements, and $SD_{start} = 10$, the standard deviation of the start value between cases.

Case re-analysis

Each constructed single-case study was re-analyzed using a randomization test for single-cases (Dugard et al., 2012). For simulations 1 to 3, a minimum phase length of five

was set for each phase, and for simulation 4 (multiple case designs), a minimum phase length of three was used. As a key statistic, we set the difference between the phase means ($M_{\text{phase B}} - M_{\text{phase A}}$). Equal phase means counted in support of the assumption of an intervention effect. A maximum of 200 permutations per analysis was created (only relevant for simulation 4). An analysis was considered to indicate a significant effect when the actual observed mean difference was equal to or higher than 95% of permutations.

Results

Phase length and level effect size (RQ 1)

Figure 3 depicts the results of simulation 1. Each orange point represents the percentage of randomization tests that yielded a significant result when a true effect was prevalent (power), while the green points indicate the alpha-error rate, which is the percentage of the randomization tests that yielded a significant result when no effect was prevalent (false positives). The two gray horizontal lines mark the conventional thresholds: 5% for alpha error and 80% for power. Panels (a to d) represent the intervention effect sizes introduced into the graphs (relevant for the orange points). The x-axis shows the length of phase A, while the shapes of the points correspond to different lengths of phase B.

For smaller effect sizes, panels a ($d =$

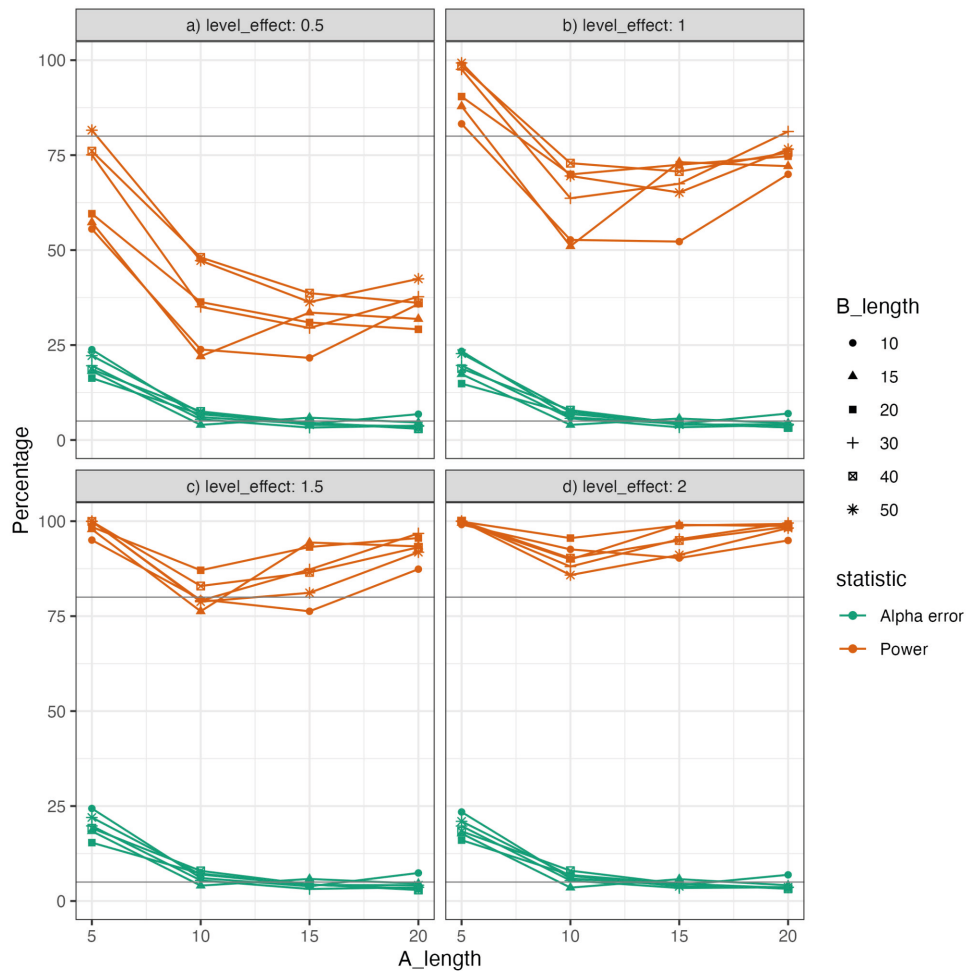


Figure 3
Power and alpha error rates (line color) for different intervention effect sizes (panel) and measurement times per phase (dot shape and x-axis; simulation 1)

0.5) and b ($d = 1.0$), the power of the randomization test is low (especially in panel a, which is mostly below 50%). Although power is higher for designs with a very short Phase A length (below 10), this comes at the cost of a highly increased alpha-error rate. For larger effect sizes, panels (c) and (d), the power is high (about or above 80%) for nearly all conditions, but, again, for very short Phase A length (below 10), this is accompanied by a high alpha-error rate. The alpha-error rate is acceptable for all conditions with a length of Phase A of 10 or higher.

Measurement reliability (RQ 2)

Figure 4 depicts the results of simulation 2. The colors and the panels (a to d) represent the same as in Figure 3. The x-axis indicates the overall number of measurements ($\frac{1}{3}$ for Phase A and $\frac{2}{3}$ for Phase B). The shapes correspond to the reliability of the measurements (measurement error). The power is low across all conditions for a smaller effect size of $d = 0.5$ (panel a). When the effect size is $d = 1.0$ (panel b), power only reaches a sufficient level (above 80%) when the reliability is very high ($r_{tt} =$

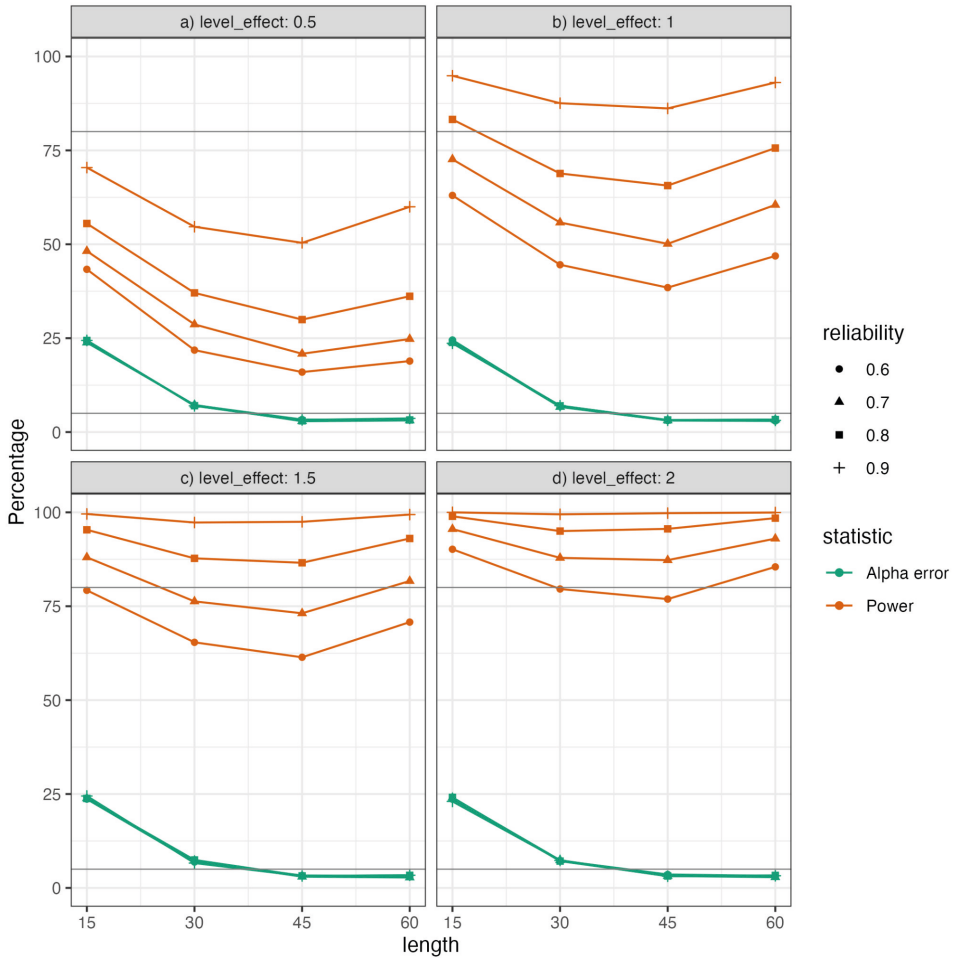


Figure 4
Power and alpha error rates (line color) for different intervention effect sizes (part), measurement times (x-axis), and measurement reliabilities (dot shape; simulation 2)

0.9). With an effect size of $d = 1.5$ (panel c), the power is consistently high for conditions where reliability is high ($r_{tt} \geq 0.8$). In addition, power exceeds 80%, even with a reliability $r_{tt} = 0.7$, when the overall number of measurements is 60. For a large level effect of $d = 2.0$ (panel d), the power is consistently at or above 80% even when the reliability $r_{tt} = 0.6$. Again, in brief designs (5 measurements in Phase A and 10 in Phase B), power is higher but comes at the cost of highly increased alpha-error rates. Across all other conditions, alpha-error remains at an acceptable level.

Baseline trend effect size (RQ 3)

Figure 5 depicts the results of simulation 3. The colors and the panels (a to d) are the same as in Figures 3 and 4. The x-axis shows the overall number of measurements ($\frac{1}{3}$ for Phase A and $\frac{2}{3}$ for Phase B), and the shapes correspond to the strength of a data trend ranging from $d = -0.5$ to $d = +0.5$ across all measurements.

The results indicate that the prevalence of a trend effect has practically no impact on power and alpha-error rate across all conditions, as indicated by the overlapping dots

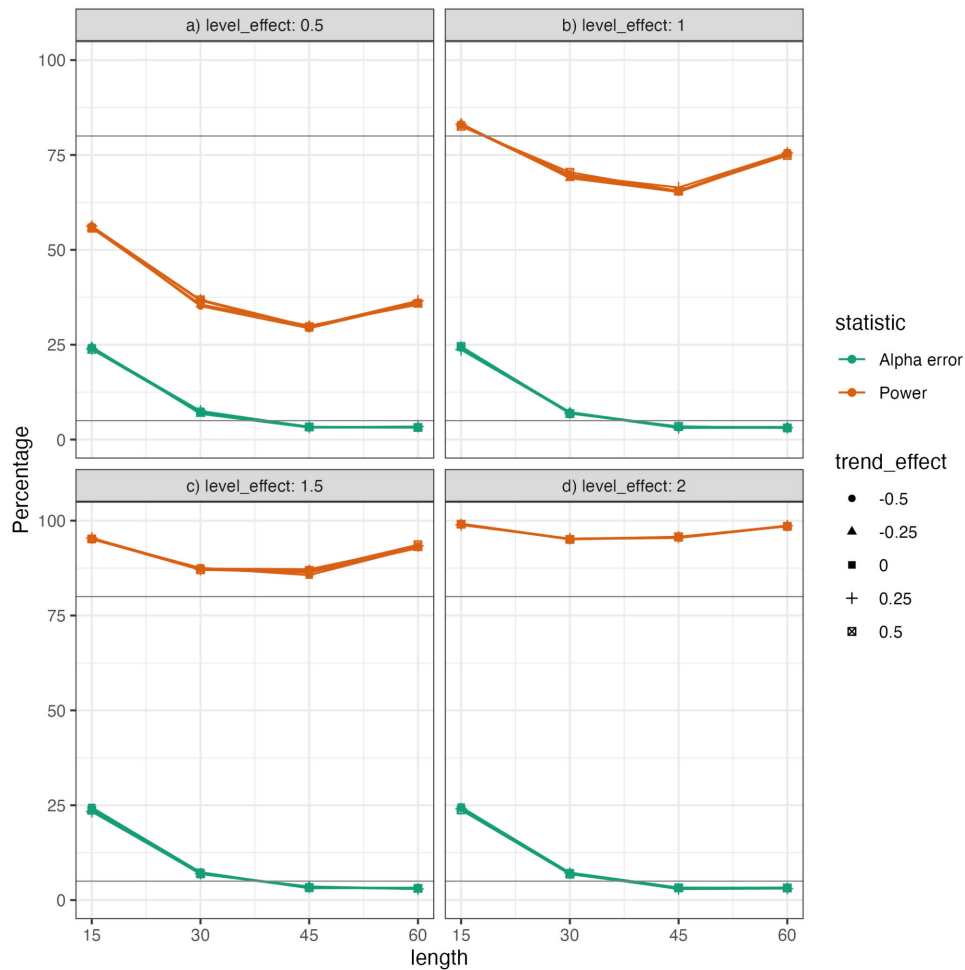


Figure 5
Power and alpha error rates (line color) for different intervention effect sizes (part), measurement times (x-axis), and different trend effect sizes (dot shape; simulation 3)

in each condition. The simulation replicates the previous findings: only effect sizes of $d \geq 1.5$ are reliably detected. For brief designs (fewer than 30 measurements), the alpha-error rate increases highly, while it remains consistently low for all other conditions.

Number of cases (RQ 4)

Figure 6 depicts the results of simulation 4. The colors represent the same as in Figures 3 to 5. The panels (a to d) depict conditions where simulated studies comprise 1, 3, 5, or 7 cases. The x-axis indicates the overall

number of measurements ($\frac{1}{3}$ for Phase A and $\frac{2}{3}$ for Phase B), and the shapes correspond to the size of the level-effect $d = 0.5$ to $d = 2.0$.

Panel (a) serves as a reference graph for designs with one case, replicating the results from Figure 3: power exceeds 80%, alpha-error remains at an acceptable rate of around 5% only when the effect size is $d \geq 1.5$, and there are at least ten measurements in Phase A and 20 measurements in Phase B (overall length = 30). For designs with 3 cases, power is above 80% for effect sizes of $d \geq 1.0$. With 5 cases, even small effects

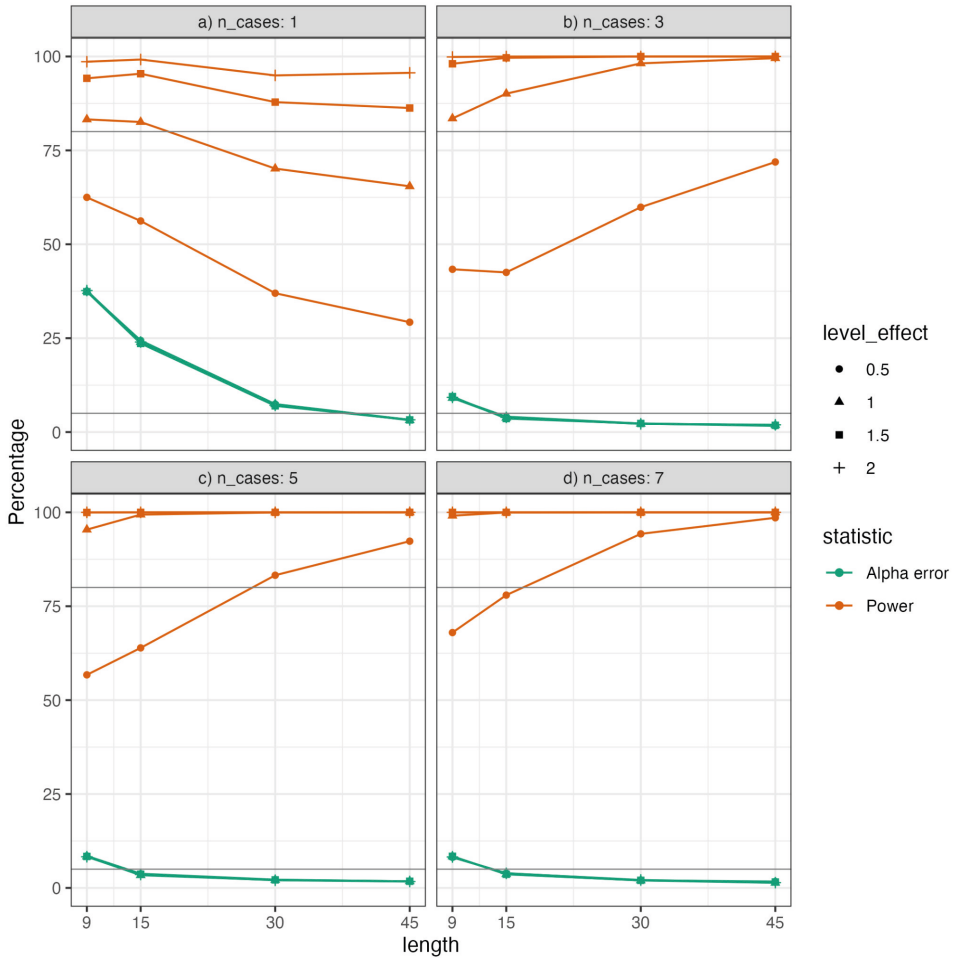


Figure 6
Power and alpha error rates (line color) for different numbers of cases per study (part), measurement times (x-axis), and various intervention effect sizes (dot shape; simulation 4)

of $d = 0.5$ can be detected with approximately 80% power when there are 30 measurements. Small effects are detected with high power in designs with seven instances when only 15 measurements are used (here, 5 in Phase A and 10 in Phase B). Even for designs with only nine measurements (3 in Phase A and 6 in Phase B), alpha-error rates remain below 10%.

Discussion

This study investigated the power and alpha-error probability when applying a randomization test for single-case studies under various design specifications. By that means, we wanted to address existing desiderata of previous studies that solely focused on isolated design parameters (e.g., auto-correlation/data trend, intervention effect size) and did not report alpha-error rates.

Overall, the results of our study demonstrate that the power of randomization tests for analyzing single-case data is highly de-

pendent on several design factors. Concerning our first research question, in single-case designs, a Phase A as well as a Phase B length of at least ten measurements is necessary to achieve an acceptable power and maintain the alpha-error at around 5%, but this is only true if measurements are reliable and the intervention effect is large ($d \geq 1.5$). Designs with a brief baseline (Phase A length < 10) have unacceptable alpha-error rates. These results extend the findings of Ferron and Onghena (1996), who showed that adequate statistical power of randomization tests in single-case becomes clear when the effect size ≥ 1.1 and phase length ≥ 5 but did not report alpha-error rates. Our implications are more conservative and align with Ferron and Ware (1995), who described inadequate statistical power of randomization tests in basic AB designs that contained 30 observations independently of the effect size and the autocorrelation of data. In any case, our results suggest that the minimum length of the baseline should exceed the lower end of design recommendations, which suggest as few as three measurements in Phase A (e.g., Kratochwill et al., 2013) and point towards a need for more intensive designs to prevent inflated alpha-error rates.

To our knowledge, no previous study examined the influence of measurement reliability on the statistical power and alpha error rates of randomization tests. Therefore, the results concerning the second research question are particularly relevant. From our findings, it becomes clear that in situations when the measurement reliability is moderate (0.7) to low (0.6), only large intervention effects ($d = 2.0$) can be detected with high power. For such large effects, the impact of reliability diminishes as power remains high. In contrast, higher measurement reliability improves power for smaller and moderate effect sizes. Researchers should, therefore, aim to include reliable measures when designing a single-case study. This in itself is challenging, as specific criteria need to be met when creating ins-

truments for repeated measures (e.g., Wilbert & Linnemann, 2011).

Concerning the third research question, it becomes clear that randomization tests are robust against baseline trends (auto-correlation), regardless of whether the trend is in the same or opposite direction of the intervention effect, with minimal impact on power and alpha-error rates. These findings contradict Ferron and Ware (1995), who indicated that the level of autocorrelation affected the power of randomization tests (even in alternating or multiple baseline designs). Similarly, Ferron and Onghena (1996) showed that negative autocorrelation 'leading to smaller powers is consistent across phase lengths and nonzero effect sizes' (p. 237) in alternating ABAB designs. Lall and Levin (2004) note "a decrease in power as the level of autocorrelation increases" (p. 76). Our findings refer to a data trend manifesting as a form of autocorrelation, such as a positive lag-1 autocorrelation, where successive measurements are positively correlated. Although this is just one type of autocorrelation, it is arguably the most critical in detecting systematic patterns over time. For this type of autocorrelation, we can confidently suggest applying randomization tests.

Finally, increasing the number of cases in multiple-baseline designs substantially enhances power and reduces the alpha-error rate. Even in very short designs (9 or 15 measurements), the power to detect an effect size of $d = 1.0$ (with 3 case designs) or $d = 0.5$ (with 7 cases) is high. A multiple-baseline design is highly recommended, as it provides good power even in small designs with moderate expected intervention effects. This finding aligns with existing design recommendations on single-case research that argue that simple AB designs do not meet evidence standards and instead recommend multiple baseline or alternating designs (Kratochwill et al., 2013).

Implications for Applied Researchers / Practice

The results of the study at hand offer insights into the statistical power of randomization tests for analyzing single-case data. They highlight requirements in terms of design length, expected intervention effect size, and the measure's reliability. When these requirements are met, randomization tests promise to yield high power at an acceptable level of alpha errors. At the same time, randomization tests require researchers to randomize facets of the design. We based our analysis on the assumption of a random intervention start-point model (Edgington, 1975), in which researchers are not completely free to decide to determine the start of a single-case intervention but instead randomly select the start point (usually from a predetermined range of possible start points) (Heyvaert & Onghena, 2014b; Levin, Ferron & Gafurov, 2014). This (restricted) randomization might conflict with a perceived pressure to change or adapt instructions/intervention as soon as possible, e.g., when baseline data indicate that students are struggling or falling behind. It is, however, crucial to mention that the validity of randomization tests always requires a kind of randomization in the design. Heyvaert and Onghena (2014b) outline results by Levin, Ferron, and Kratochwill (2012) who compared "simulation results for the application of RT [randomization test] procedures in non-randomized designs to the application in randomized designs" and found that in "former designs the test often failed to have Type I error rates that matched the nominal level, while for the latter designs Type I error rates were well controlled" (p. 60). The levels of design randomization might differ depending on the basic single-case design applied. The basic intervention start-point model has been extended over the years (Levin et al., 2014). Consequently, different randomization strategies exist for multiple baseline designs (e.g., Mara-

scuilo-Busk, Wampold-Worsham, Koehler-Levin; see Heyvaert & Onghena, 2014b).

Although the figures and data in this paper can help guide decisions regarding the design of a single-case study, their utility is limited by the multitude of design requirements. Therefore, we recommend conducting a power analysis before determining the final parameters of a planned single-case study, as this aligns with best practices in group-based research methods. A comprehensive software solution for performing this analysis is provided in the R package *scan* (Wilbert & Lüke, 2024).

Limitations

The present study has major limitations that affect the interpretation of our findings. Our simulated designs incorporated a random intervention start-point model and might not be comparable to other simulation studies based on different randomization strategies. In addition, the randomization tests applied focus on the mean difference between phases. In theory and practice, different randomization statistics are being applied (e.g., differences of medians, regression slopes, or Hedges' *g*). Incorporating different statistics might lead to differing results.

Moreover, we only focused on some design characteristics. Additional design parameters that influence the power and alpha-error rates might be incorporated in future research, e.g., data outliers/extreme points. We focused on intervention effects that result in a mean change between phases. However, intervention effects in single-case designs may comprise both level and slope effects.

It is important to note that the appropriate choice of the test statistic in a randomization test (e.g., a mean difference or a slope difference) depends not only on the statistical power of the test, but crucially on the effects that are expected *a priori*. In other words, the results of a randomization test should be interpreted as a test of the researchers' hypotheses. There is a risk that

researchers may be tempted to run multiple randomization tests using different test statistics and, based on the results, decide *a posteriori* which process best explains the data.

It is also essential to consider to what extent our results apply to real-world data, as the analyses were based on simulated data. Real-world data may not adhere to the distributional assumptions of our simulations and could be more susceptible to outliers.

References

- Briesch, A. M., & Briesch, J. M. (2016). Meta-Analysis of Behavioral Self-Management Interventions in Single-Case Research. *School Psychology Review*, 45(1), 3–18. <https://doi.org/10.17105/spr45-1.3-18>
- Edgington, E. S. (1975). Randomization Tests for One-Subject Operant Experiments. *The Journal of Psychology*, 90(1), 57–68. <https://doi.org/10.1080/00223980.1975.9923926>
- Edgington, E., & Onghena, P. (2007). *Randomization Tests*, 4th Edn. Boca Raton, FL: CRC Press.
- Ferron, J., & Onghena, P. (1996). The Power of Randomization Tests for Single-Case Phase Designs. *The Journal of Experimental Education*, 64(3), 231–239. <https://doi.org/10.1080/00220973.1996.9943805>
- Ferron, J., & Sentovich, C. (2002). Statistical Power of Randomization Tests Used with Multiple-Baseline Designs. *The Journal of Experimental Education*, 70(2), 165–178. <https://doi.org/10.1080/00220970209599504>
- Ferron, J., & Ware, W. (1995). Analyzing Single-Case Data: The Power of Randomization Tests. *The Journal of Experimental Education*, 63(2), 167–178. <https://doi.org/10.1080/00220973.1995.9943820>
- Herzog, M., & Casale, G. (2024). Inattention negatively moderates the effectiveness of a mathematics intervention in low performing primary school students. *Frontiers in Education*, 9, 1276741. <https://doi.org/10/gt49vn>
- Heyvaert, M., & Onghena, P. (2014a). Analysis of single-case data: Randomisation tests for measures of effect size. *Neuropsychological Rehabilitation*, 24(3–4), 507–527. <https://doi.org/10.1080/09602011.2013.818564>
- Heyvaert, M., & Onghena, P. (2014b). Randomization tests for single-case experiments: State of the art, state of the science, and state of the application. *Journal of Contextual Behavioral Science*, 3(1), 51–64. <https://doi.org/10.1016/j.jcbs.2013.10.002>
- Ma, H.-H. (2006). An Alternative Method for Quantitative Synthesis of Single-Subject Researches: Percentage of Data Points Exceeding the Median. *Behavior Modification*, 30(5), 598–617. <https://doi.org/10.1177/0145445504272974>
- Klapproth, F. (2018). Biased predictions of students' future achievement: An experimental study on pre-service teachers' interpretation of curriculum-based measurement graphs. *Studies in Educational Evaluation*, 59, 67–75. <https://doi.org/10.1016/j.stueduc.2018.03.004>
- Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., et al. (2013). Single-case intervention research design standards. *Remedial Spec. Educ.* 34, 26–38. doi: 10.1177/074193251245 2794
- Lall, V. F., & Levin, J. R. (2004). An empirical investigation of the statistical properties of generalized single-case randomization tests. *Journal of School Psychology*, 42(1), 61–86. <https://doi.org/10.1016/j.jsp.2003.11.002>
- Ledford, J. R., Lambert, J. M., Pustejovsky, J. E., Zimmerman, K. N., Hollins, N., & Barton, E. E. (2022). Single-Case-Design Research in Special Education: Next-Generation Guidelines and Considerations. *Exceptional Children*, 001440292211376. <https://doi.org/10.1177/00144029221137656>
- Levin, J. R., Ferron, J. M., & Kratochwill, T. R. (2012). Nonparametric statistical tests for single-case systematic and randomized ABAB...AB and alternating treatment intervention designs: New developments, new directions. *Journal of School Psychology*, 50(5), 599–624. <https://doi.org/10.1016/j.jsp.2012.05.001>
- Levin, J. R., Ferron, J. M., & Gafurov, B. S. (2014). Improved Randomization Tests for a Class of Single-Case Intervention Designs. *Journal of Modern Applied Statistical Methods*, 13(2), 2–52. <https://doi.org/10.22237/jmasm/1414814460>
- Manolov, R., & Moeyaert, M. (2017). How Can Single-Case Data Be Analyzed? Software Resources, Tutorial, and Reflections on Analysis. *Behavior Modification*, 41(2), 179–228. <https://doi.org/10.1177/0145445516664307>

- Manolov, R., & Solanas, A. (2009). Problems of the randomization test for AB designs. *Psicológica*, 30, 137–154.
- Manolov, R., Solanas, A., Bulté, I., & Onghe-
na, P. (2009). Data-Division-Specific Robust-
ness and Power of Randomization Tests for
ABAB Designs. *The Journal of Experimen-
tal Education*, 78(2), 191–214. [https://doi.
org/10.1080/00220970903292827](https://doi.org/10.1080/00220970903292827)
- Parker, R. I., Vannest, K. J., & Davis, J. L. (2011).
Effect Size in Single-Case Research: A Review of
Nine Nonoverlap Techniques. *Behavior Modifi-
cation*, 35(4), 303–322. <https://doi.org/10/dsdfs4>
- Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S.
B. (2011). Combining Nonoverlap and Trend for
Single-Case Research: Tau-U. *Behavior Therapy*,
42(2), 284–299. <https://doi.org/10/fxcfr>
- Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987).
The Quantitative Synthesis of Single-Subject Re-
search: Methodology and Validation. *Remedial
and Special Education*, 8(2), 24–33. [https://doi.
org/10.1177/074193258700800206](https://doi.org/10.1177/074193258700800206)
- Soares, D. A., Harrison, J. R., Vannest, K. J., &
McClelland, S. S. (2016). Effect Size for Token
Economy Use in Contemporary Classroom Set-
tings: A Meta-Analysis of Single-Case Research.
School Psychology Review, 45(4), 379–399.
<https://doi.org/10.17105/spr45-4.379-399>
- Sperling, M., Barwasser, A., & Grünke, M. (2019).
*The Effects of a Reading Racetrack Intervention
on the Sight Word Fluency of Learning Disabled
Elementary School Students With German as Se-
cond Language*.
- Wilbert, J., Börnert-Ringleb, M., & Lüke, T. (2022).
Statistical Power of Piecewise Regression Analy-
ses of Single-Case Experimental Studies Address-
ing Behavior Problems. *Frontiers in Education*.
<https://doi.org/10.3389/feduc.2022.917944>
- Wilbert, J., & Linnemann, M. (2011). Kriterien zur
Analyse eines Tests zur Lernverlaufdiagnostik.
Empirische Sonderpädagogik, 3, 225–242.
- Wilbert, J., & Lüke, T. (2024). *Scan: Single-Case Data
Analyses for Single and Multiple Baseline Designs*.
(Version 0.61.1) [English]. University of Münster.
<https://CRAN.R-project.org/package=scan>

Author Note

 Jürgen Wilbert
<https://orcid.org/0000-0002-8392-2873>

 Moritz Börnert-Ringleb
<https://orcid.org/0000-0003-3533-0993>

 Timo Lüke
<https://orcid.org/0000-0002-2603-7341>

Corresponding author

Jürgen Wilbert
University of Münster
Georgskommende 26
D 308, D-48143 Münster
juergen.wilbert@uni-muenster.de

Supplement

<https://doi.org/10.17605/OSF.IO/GMY26>

Erstmals eingereicht: 22.10.2024
Überarbeitung eingereicht: 31.03.2025
Angenommen: 01.04.2025

Offene Daten	Der entsprechende R-Code ist unter https://doi.org/10.17605/OSF.IO/GMY26 verfügbar.
Offener Code	Der entsprechende R-Code ist unter https://doi.org/10.17605/OSF.IO/GMY26 verfügbar.
Offene Materialien	/
Präregistrierung	Nein
Votum Ethikkommission	Nein
Finanzielle und weitere sachliche Unterstützung	Es erfolgte keine finanzielle und weitere sachliche Unterstützungen durch Dritte.
Autorenschaft	JW did the conceptualization, carried out the data curation, formal analysis, and software, investigated the data, performed the methodology, visualized the data, wrote the original draft, and wrote, reviewed, and edited the manuscript. MB-R and TL did the conceptualization, investigated the data, performed the methodology, wrote the original draft, and wrote, reviewed, and edited the manuscript. All authors contributed to the article and approved the submitted version.
KI und generative Modelle	DeepL was used for language editing.

Heterogenität im Individuum: die Bedeutung von intraindividueller Variabilität aus der Perspektive der Sonderpädagogik

Julia Winkes

Universität Freiburg

Zusammenfassung

Dieser Beitrag beschreibt intraindividuelle Variabilität (IIV) als bedeutsame Informationsquelle im sonderpädagogischen Kontext. Als IIV werden Schwankungen in Leistung und Verhalten einer Person über verschiedene Zeitpunkte verstanden. In der Sonderpädagogik wird IIV häufig als Messfehler interpretiert, was die Bedeutung dieser Dynamiken unterschätzt, denn sie können wertvolle Einblicke in kognitive Prozesse, Lernen und Entwicklung ermöglichen. Der Beitrag beleuchtet zwei zentrale Perspektivenwechsel: Erstens wird die Bedeutung von IIV über Mittelwertanalysen hinaus anerkannt, indem IIV als eigenständige Informationsquelle betrachtet wird, was am Beispiel von Resultaten zu IIV in Reaktionszeit- und Arbeitsgedächtnisaufgaben verdeutlicht wird. Zweitens wird der Fokus von Gruppenergebnissen zu individuellen Entwicklungsprozessen verschoben, da psychologische Prozesse oft nicht-ergodisch sind und daher gruppenbasierte Ergebnisse nicht auf Einzelpersonen übertragbar sind. Diese Sichtweise wird anhand von Complex Dynamic Systems Theorie als theoretischer Bezugspunkt und Experience Sampling Methods (ESM) als methodischem Zugang verdeutlicht. Die Implikationen dieser Erkenntnisse für die Sonderpädagogik umfassen verschiedene Aspekte in den Bereichen Diagnostik, Förderung und Forschung. Während in der Diagnostik eine Sensibilisierung auf potenzielle Schwierigkeiten durch IIV notwendig ist, eröffnen sich für die sonderpädagogische Förderung innovative Potenziale in Bezug auf die adaptive Anpassung von Interventionsstrategien an individuelle Lernverläufe. In der Forschung kann die Sonderpädagogik auf der einen Seite durch Methoden wie die Lernverlaufsdiagnostik zur Erweiterung des Kenntnisstandes beitragen. Auf der anderen Seite steht mit ESM ein starkes Werkzeug zur Verfügung, um Variabilität in dynamischen Kontexten des täglichen Lebens auch bei Personenkreisen mit Förderbedarfen zu untersuchen.

Schlagwörter: Intraindividuelle Variabilität, Diagnostik, Förderung, Forschung

Heterogeneity Within the Individual: The Role of Intraindividual Variability in Special Education

Summary

This article describes intra-individual variability (IIV) as a significant source of information in the context of special education. IIV refers to fluctuations in an individual's performance

and behavior over various time points. In special education, IIV is often interpreted as measurement error, which underestimates the importance of these dynamics, as they can provide valuable insights into cognitive processes, learning, and development. The article highlights two key shifts in perspective: First, it recognizes the importance of IIV beyond mean-level analyses by considering IIV as an independent source of information, illustrated through findings on IIV in reaction time and working memory tasks. Second, it shifts the focus from group results to individual developmental processes, since psychological processes are often non-ergodic, making group-based results non-transferable to individuals. This perspective is explained using Complex Dynamic Systems Theory as a theoretical framework and Experience Sampling Methods (ESM) as a methodological approach. The implications of these insights for special education encompass various aspects in assessment, intervention, and research. In assessment, it is essential to increase awareness of challenges posed by IIV. In the area of special needs education, innovative potential is opening up for adaptively tailoring intervention strategies to individual learning trajectories. In research, special education can, on the one hand, contribute to the expansion of knowledge through methods such as curriculum-based measurement. On the other hand, ESM provides a powerful tool for investigating variability in dynamic contexts of everyday life, including in groups of people with special educational needs.

Keywords: Intraindividual Variability, assessment, intervention, research

Schwankungen in Leistung und Verhalten sind für sonderpädagogische Fachpersonen wohlbekannte Phänomene und integraler Bestandteil ihres Arbeitsalltags. Ein Kind zeigt an einem Tag eine unerwartet gute oder schwache Leistung im Lesen oder in der Mathematik; Verhaltensprobleme treten mal in größerem oder in geringerem Ausmaß auf. Obgleich die Feststellung, dass menschliches Verhalten instabil ist und von Situation zu Situation variiert, weder überraschend noch neu ist, sind diese Fluktuationen in der Sonderpädagogik selten selber Gegenstand der Reflexion in Forschung und Praxis. In anderen Disziplinen, wie etwa in der Entwicklungspsychologie, gibt es unterdessen wenig Zweifel daran, dass die Untersuchung von intraindividueller Variabilität (IIV), also der Variation innerhalb einer Person zwischen verschiedenen (Mess-)zeitpunkten (van Dijk & van Geert, 2015), Einblicke in kognitive Verarbeitungsprozesse, Lernen und Entwicklung gibt und daher eine fruchtbare Informationsquelle darstellen kann (Nesselroade & Molenaar, 2010).

Aufbauend auf dieser Erkenntnis diskutiert der vorliegende Beitrag die Relevanz von IIV im sonderpädagogischen Kontext. Er zeichnet zwei zentrale Perspektivenwechsel nach: Der erste thematisiert die Bedeutung von IIV über Mittelwertanalysen hinaus. Schwerpunktmäßig werden Befunde zu intraindividuellen Schwankungen in kognitiven Grundfunktionen, insbesondere in Arbeitsgedächtnis- und Reaktionszeitaufgaben, dargestellt. Dabei wird aufgezeigt, bei welchen Personengruppen eine erhöhte IIV auftritt und welche funktionale Bedeutung diesem Befund auf Aufmerksamkeitskontrolle und Informationsverarbeitung zugeschrieben wird. Zudem werden Zusammenhänge zwischen IIV, schulischen Leistungen und Alltagsverhalten erläutert. Der zweite Perspektivenwechsel skizziert eine Abkehr von der traditionellen Annahme stabiler Personenmerkmale. Er geht davon aus, dass psychologische Prozesse häufig nicht-ergodisch sind – das heißt, dass Variabilität innerhalb von Individuen (IIV) eigenständige Entwicklungen widerspiegelt und nicht einfach aus Durchschnittswerten

oder Gruppenbefunden erschlossen werden kann (Molenaar, 2004). Diese Sichtweise wird theoretisch durch die Complex Dynamic Systems Theory (CDST) fundiert, die Entwicklung als dynamischen, nichtlinearen Prozess versteht. Methodisch wird mit den Experience Sampling Methods (ESM) ein Ansatz vorgestellt, der intraindividuelle Schwankungen im Alltag sichtbar und analysierbar macht. Im abschließenden Abschnitt werden die Implikationen der Perspektivenwechsel für die sonderpädagogische Diagnostik, Förderung und Forschung aufgezeigt: die Bedeutung von IIV für diagnostische Einschätzungen, Ansätze zur stärkeren Individualisierung von Fördermaßnahmen sowie erste Perspektiven für eine intensivere Untersuchung intraindividuellen Dynamiken in der sonderpädagogischen Forschung.

Perspektivenwechsel 1: Intraindividuelle Variabilität besteht nicht (nur) aus Messfehlern

Es gibt in der Sonderpädagogik, ebenso wie in vielen anderen Disziplinen, eine lange Tradition, Fluktuationen im Verhalten von Personen als Messfehler zu betrachten und durch Aggregieren und *data smoothing* eine Näherung an den «wahren Wert» einer Person erreichen zu wollen (Lievens et al., 2018). Der Orientierungspunkt ist damit das einflussreiche Reliabilitätskonzept der klassischen Testtheorie, das Variabilität innerhalb einer Person in der Regel entweder als Messfehler behandelt, der aus unzureichenden Instrumenten resultiert, oder durch Änderungen im Kontext hervorgerufen sieht (van Dijk & van Geert, 2015; van Geert & van Dijk, 2002). IIV wird in diesem Verständnis gleichgesetzt mit Fehlervarianz, die es so gut wie möglich methodisch auszuschalten gilt, da nur der wahre Wert einer Person über Aussagekraft verfügt (Lievens et al., 2018). Auch statistische Verfahren, die Fehlervarianz systematisch in verschiedene Komponenten unterteilen (z.B. Generalisierbarkeitstheorie, Mehrebenenmodelle,

Item-Response-Theorie), zielen in ihrer klassischen Anwendung nicht darauf ab, intraindividuelle Fluktuationen als eigenständige Informationsquelle zu interpretieren. Ohne die Existenz von Messfehlern in Frage zu stellen, verbreitet sich in neuerer Zeit mehr und mehr eine erweiterte Perspektive auf IIV, deren wichtigste Annahmen Schmiedek et al. (2009) folgendermaßen zusammenfassen: «Intraindividual fluctuations in behavior may (a) be separable from measurement error, (b) differ in magnitude between individuals, and (c) predict other psychological phenomena” (S. 841). IIV wird damit als distinkte und bedeutungsvolle Information betrachtet, die zusätzlich zum Mittelwert analysiert werden kann und selber eine prädiktive Funktion einnimmt. Ihre Analyse ist verbunden mit der Annahme, die Dynamik von kognitiven Prozessen und von Entwicklung und Lernen besser zu verstehen (Amenta & Crepaldi, 2016). Ein bedeutsamer Anteil der Forschung mit Fokus auf intraindividuellen Schwankungen befasst sich mit Variabilität in wichtigen kognitiven Grundfunktionen, insbesondere Reaktionszeit und Arbeitsgedächtnis. Solche Aufgaben können auf verschiedenen Zeitskalen eingesetzt werden: von Tag zu Tag, mehrfach innerhalb eines Tages oder sogar mehrfach innerhalb eines Versuchs (Moment-zu-Moment-Ebene; Dirk & Schmiedek, 2016). Variabilität kann es auf allen möglichen Zeitskalen geben. Der Begriff wird aber am häufigsten für Schwankungen in einem relativ kurzen Zeitrahmen verwendet (Van Dijk und van Geert (215). Je spezifischer die Aufgabe ist, desto häufiger und genauer sollte das Verhalten im Zeitverlauf erhoben werden. Für allgemeine Fähigkeiten, die sich langsam entwickeln, können die zeitlichen Abstände zwischen den Messungen entsprechend ausgeweitet werden (Verspoor & de Bot, 2021).

Ogleich IIV immer auch von situationalen Komponenten wie Stress, Schlaf oder Lärm abhängig ist (Dirk & Schmiedek, 2017; Galeano-Keiner et al., 2022), wird das typische Level von IIV innerhalb einer

Person in Bezug auf die kognitiven Grundfunktionen als relativ stabil beschrieben und es gibt somit Personen, die mehr IIV aufweisen als andere (Ali et al., 2018). Über die Lebensspanne hinweg entwickelt sich IIV in Reaktionszeitaufgaben im Sinne einer U-Kurve: In der frühen Kindheit ist IIV typischerweise hoch, sie verringert sich im Verlaufe der Kindheit und in der Adoleszenz und beginnt im späteren Erwachsenenalter wieder anzusteigen (Cubillo et al., 2023; Fagot et al., 2018). Neben Unterschieden, die mit der Entwicklungsphase der Person in Zusammenhang stehen, ist eine erhöhte IIV bei verschiedenen klinischen Störungsbildern gut dokumentiert, im Erwachsenenalter z.B. bei Demenz, Schizophrenie oder bei traumatischen Hirnverletzungen (siehe Cubillo et al., 2023). Im sonderpädagogischen Kontext von hoher Relevanz sind Befunde zu diversen Entwicklungsstörungen. Allen voran wird IIV sehr konsistent als typisches Defizit von Personen mit Aufmerksamkeits-Defizit-Hyperaktivitätsstörung (ADHD) beschrieben (Ali et al., 2018; Borella et al., 2011; Gooch et al., 2012) und sogar als «fundamental marker of ADHD» bezeichnet (Gooch et al., 2012, S. 468). In der Zwischenzeit gibt es zunehmend Hinweise darauf, dass eine erhöhte IIV nicht spezifisch für ADHD ist, sondern auch bei anderen Entwicklungsstörungen, wie Autismus-Spektrum-Störung (ASS; Geurts et al., 2008), Fetale Alkohol Spektrum Störung (FASD; Ali et al., 2018) und Dyslexie (Borella et al., 2011) auftritt. Dirk und Schmiedek (2016) berichten, dass Kinder mit einer niedrigeren fluiden Intelligenz eine höhere Variabilität in Arbeitsgedächtnisaufgaben von Tag zu Tag aufweisen. Einige Autoren betrachten eine erhöhte IIV in Reaktionszeitaufgaben daher eher als Anzeichen für eine unspezifische pathophysiologische Verarbeitung (Castellanos et al., 2005) und weniger als spezifisches Merkmal von ADHD.

Auswirkungen von IIV in kognitiven Funktionen auf schulisches Lernen und Verhalten

Wie im vorhergehenden Abschnitt dargestellt wurde, wird IIV in kognitiven Grundfunktionen allgemein interpretiert als Marker für Schwierigkeiten in der Informationsverarbeitung, die sich auf die Aufrechterhaltung der Aufmerksamkeit und auf die exekutive Kontrolle auswirken (Borella et al., 2011; Isbell et al., 2018). Eine geringere IIV wird daher – spezifisch im Kontext von diesen Aufgaben – als positiv bewertet und als Signal für Stabilität und Effizienz von kognitiven Prozessen ausgelegt. Diese Interpretation wird unterstrichen durch Studien, in denen ein adaptives Arbeitsgedächtnistraining zu einer geringeren post-Training IIV führte und eine geringere IIV mit einer besseren Aufgabenperformanz einherging (Cañigual et al., 2023; Cubillo et al., 2023). Diese Resultate sind bereits für sich interessant, allerdings sind aus sonderpädagogischer Perspektive Auswirkungen dieser Schwankungen auf das Verhalten von Personen von größerer praktischer Relevanz. Da IIV in Reaktionszeitaufgaben als Index für die Aufmerksamkeitskontrolle fungiert, ist davon auszugehen, dass Fluktuationen von Aufmerksamkeitsressourcen zu beeinträchtigten Lernprozessen führen. Kautto et al. (2024) vermuten spezifisch für die Sprachentwicklung ein Defizit bei der effizienten Nutzung des Inputs, welches das sprachliche Lernen beeinträchtigt: «This could be compared to listening in noisy environment but the «noise» comes from neurocognitive processes themselves, yielding to low signal-to-noise-ratio» (S. 2704). Tatsächlich untersuchen mehrere Forschungsarbeiten den Zusammenhang zwischen IIV in kognitiven Aufgaben (Reaktionszeit, Arbeitsgedächtnis) und den aktuellen und zukünftigen Schulleistungen von Kindern. Isbell et al. (2018) erhoben IIV in Reaktionszeitaufgaben bei Vorschulkindern und konnten zeigen, dass Kinder mit einer höheren IIV schlechtere Schulleistungsbedingungen

aufwiesen, was wiederum ein Prädiktor für die Leistungen in der ersten Klasse darstellte. Bei Judd et al. (2021) erwies sich IIV im Arbeitsgedächtnis bei Sechsjährigen als prädiktiv für die mathematische Kompetenz drei Jahre später. Einen Zusammenhang zwischen IIV und mathematischer Entwicklung konnten Wang et al. (2022) nicht feststellen, wohl aber zum sprachlichen Lernen bei chinesischen Primarschulkindern. Auch bei Kautto et al. (2024) war eine höhere IIV in einer visuellen Reaktionszeitaufgabe assoziiert mit schwächeren sprachlichen Leistungen.

Mit täglichem Verhalten kann IIV in kognitiven Grundfunktionen ebenfalls in Verbindung gebracht werden. Beispielsweise zeigen Kinder und Jugendliche mit einer höheren IIV in Reaktionszeitaufgaben häufiger externalisierende Verhaltensprobleme und weniger soziale Kompetenzen. Zur Erklärung dieses Zusammenhangs wird aufgeführt, dass eine fluktuierende Aufmerksamkeit sich negativ auf die Verarbeitung sozialer Informationen auswirkt, beispielsweise auf die Wahrnehmung und Interpretation von sozialen Cues (Tamm et al., 2019). Zeng et al. (2022) weisen darauf hin, dass eine erhöhte IIV in Reaktionszeitaufgaben im Kindesalter einen Indikator für unausgereifte exekutive Hirnfunktionen darstellt (vgl. die oben beschriebene U-Funktion in der Entwicklung über die Lebensspanne), die wiederum Impulsunterdrückung und Selbstkontrolle in stressbehafteten Situationen negativ beeinträchtigen. Bei älteren Erwachsenen konnte ein Zusammenhang mit Problemlösefähigkeiten aufgezeigt werden, der auf eine funktionale Relevanz von IIV für die Bewältigung des täglichen Lebens hindeutet (Burton et al., 2009). Substanzielle intraindividuelle Unterschiede in der Selbstregulation existieren bei Kindern verschiedenen Alters (Blume et al., 2022; Ludwig et al., 2016) und es kann angenommen werden, dass Fluktuationen in den exekutiven Funktionen sich folglich nicht nur auf das schulische Lernen auswirken, sondern auch mit sozialen und lebenspraktischen

Fähigkeiten assoziiert sind (z.B. Wolfe et al., 2013).

Am Beispiel von IIV in kognitiven Grundfunktionen konnte aufgezeigt werden, dass die Analyse von intraindividuellen Schwankungen relevante Informationen bereitstellt, die über das durchschnittliche Leistungsniveau hinaus prädiktiven Wert für andere Phänomene besitzt. Studien zu dieser Thematik umfassen in der Regel Gruppenvergleiche, also z.B. den Vergleich von jüngeren und älteren Personen oder von Personen mit und ohne Entwicklungsstörungen oder kognitiven Schwierigkeiten. Ein weiterer Perspektivenwechsel rückt den Fokus weg von Mittelwertvergleichen zwischen Gruppen hin zur Analyse von Variabilität innerhalb von Personen über die Zeit.

Perspektivenwechsel 2: Psychologische Prozesse sind nicht-ergodisch

Das ergodische Prinzip besagt, dass Individuen sich gleich verhalten wie die Gruppe und dass die untersuchten Prozesse über die Zeit im Individuum stabil sind. Tatsächlich trifft das in vielen Bereichen nicht zu und Beobachtungen in Bezug auf interindividuelle Unterschiede können nicht auf ein Individuum angewendet werden, um gleiche Schlussfolgerungen über die intraindividuelle Variabilität zu ziehen (Molenaar, 2004). Damit lassen sich Erkenntnisse aus Gruppenstudien, z.B. zum Lernen, Verhalten oder Erleben nicht auf Individuen generalisieren. Stattdessen existiert eine beträchtliche Heterogenität sowohl zwischen Personen, als auch innerhalb von Personen über die Zeit (Molenaar & Campbell, 2009; Saqr, 2024). Nomothetische Ansätze, die auf Basis von Gruppenstudien generelle Vorhersagen über eine Population machen, sollten daher ergänzt werden durch idiographische statistische Auswertungen. Diese untersuchen Variabilität innerhalb von Personen und benötigen daher intensive Datenerhebungsmethoden über verschiedene Zeitpunkte (Beltz et al., 2016; van Dijk et al., 2024). Im Folgenden werden

zwei Ansätze vorgestellt, für welche die Abkehr vom ergodischen Prinzip ein wesentliches Kennzeichen darstellt. Dabei handelt es sich zum einen um Complex Dynamic Systems Theory (CDST), die einen theoretischen Bezugspunkt bietet und um Experience Sampling Methods (ESM) als methodische Herangehensweise.

Complex Dynamic Systems Theory

CDST stellt eine Metatheorie dar, die für verschiedenste Bereiche menschlichen Verhaltens anwendbar ist, in den vergangenen Jahren aber insbesondere Einfluss in der Mehrsprachigkeitsforschung erlangt hat. Grundlegend ist die Idee von Entwicklung als einem selbstorganisierenden System, das aus verschiedenen Subsystemen besteht (Lowie & Verspoor, 2019). Diese Komponenten (z.B. Wahrnehmung, Kognition, Gedächtnis, Motivation) interagieren dynamisch miteinander und mit dem sozialen und materiellen Kontext (z.B. der Aufgabe, dem sprachlichen Input). Während dieses Austausches zwischen den Komponenten verändern und formen sich diese gegenseitig. Unterschieden wird zwischen relativ stabilen Phasen (attractor states) und Transitionsphasen, in denen Entwicklung vor allem stattfindet. IIV in den entsprechenden Komponenten tritt in beiden Phasen auf, in den Transitionsphasen aber in erhöhtem Ausmass. Sie wird als Anzeichen dafür betrachtet, dass neue Strategien und Fähigkeiten exploriert werden. Diese Versuche sind in dem Moment nicht immer erfolgreich, führen aber im Endeffekt zum Erwerb von neuen Kompetenzen, oder wie Lowie und Verspoor (2019, S. 191) festhalten: "Systems have to become unstable before they can change." Damit zeigt sich auch, dass IIV keine universelle Bedeutung hat, sondern dass diese abhängig ist von der untersuchten Aufgabe und von der Entwicklungsphase, in der eine Person sich befindet (Diehl et al., 2015): Während die oben referierten Resultate aus Studien zu kognitiven Basisprozessen IIV als Vul-

nerabilitätsfaktor betrachten und negativ konnotieren, bezeichnet CDST IIV als notwendigen Vorläufer von Entwicklung. Die scheinbaren Widersprüche werden in der Regel durch die unterschiedlichen Aufgabenstellungen erklärt: Bei Aufgaben, die mit Lernen, höheren kognitiven Anforderungen oder dem Explorieren von neuen Strategien verbunden sind, scheint eine höhere IIV ein positiver Indikator zu sein, der mit höheren Leistungen und einer schnelleren Entwicklung in Zusammenhang steht, was z.B. für die mehrsprachige Sprachentwicklung auch in verschiedenen Studien nachgewiesen werden konnte (Huang et al., 2021; Lowie & Verspoor, 2019). Aber auch innerhalb von kognitiven Aufgabenstellungen zeigten Cañigual et al. (2023) jüngst, dass bei Aufgabenstellungen, bei denen eine größere Flexibilität zu einem besseren Ergebnis führt, eine erhöhte IIV als adaptiv zu bewerten ist.

Eine zentrale Annahme von CDST besteht darin, dass Entwicklung durch idiosynkratische Pfade und nonlineare Dynamiken gekennzeichnet ist (Verspoor & de Bot, 2021). Vertreter der CDST erwarten, dass eine erhöhte IIV auf der individuellen Ebene ein Prädiktor für anschließende Lern- und Entwicklungsfortschritte dieses Individuums darstellt. Um diese Mechanismen zu beobachten, müssen einzelne Individuen intensiv über längere Zeiträume untersucht werden, denn bei keinen zwei Personen werden diese individuellen Entwicklungsdynamiken mit dem Wechsel zwischen stabilen Phasen und Transitionsphasen in exakt gleicher Weise ablaufen. Das typische Forschungsdesign besteht in der Tradition von CDST daher aus Einzelfallstudien mit möglichst dichten wiederholten Messungen über die Zeit mit einem Fokus auf Interaktionen zwischen verschiedenen Variablen (Komponenten des Systems) und auf der Existenz von IIV, die über die Zeit selber variiert (van Dijk et al., 2024; Verspoor & de Bot, 2021). Die bevorzugte Wahl von Einzelfalldesigns ist teilweise aber auch dadurch begründet, dass die engmaschige Erhebung von Daten

innerhalb von Individuen über längere Zeiträume für CDST-inspirierte Studien Forscher vor massive Schwierigkeiten in Bezug auf die Durchführbarkeit ihrer Forschungsvorhaben stellt (Bulté & Housen, 2020; van Geert & van Dijk, 2002). Starke methodische Tools für die Untersuchung von Variabilität innerhalb von Personen im Sinne von Mess- und Modellierungstechniken wurden in jüngerer Zeit im Zusammenhang mit Intensive Longitudinal Methods entwickelt (Beck & Jackson, 2021), die daher im folgenden Abschnitt angesprochen werden.

Experience Sampling Methods

Experience Sampling Methods (ESM), bei denen Personen mehrfach im Alltag befragt werden, um momentane Erlebnisse und Zustände möglichst unmittelbar zu erfassen, ermöglichen eine dichte, zeiteffiziente und valide Erhebung von Merkmalen, die über kurze Zeiträume intraindividuell stark schwanken; dabei kommen häufig mobile Endgeräte wie Smartphones zum Einsatz, die flexible und zeitnahe Datenerhebungen im natürlichen Lebenskontext erlauben. Die Auswertung von Verläufen und Prozessen innerhalb eines Individuums sind mit dieser Methodik ebenso indiziert wie Vergleiche zwischen Individuen, weshalb die Analysen durch Multi-Level-Modelle erfolgen (Könen & Karbach, 2015; Venetz & Zurbiggen, 2015), die z.T. auch explizit dafür konzipiert sind, nomothetische und idiographische Analysen miteinander zu verbinden (z.B. GIMME, siehe Beltz et al., 2016). Der Mehrwert von ESM-Methoden soll hier beispielhaft an Studien zu Emotions- und Selbstregulationsstrategien verdeutlicht werden, da diese z.B. bei Personen mit emotionalen Störungen und mit Verhaltensauffälligkeiten auch für die Sonderpädagogik von Interesse sind. Traditionelle Forschung in diesem Bereich erfasst Strategiekennntnis und -gebrauch zu einem einzelnen Zeitpunkt (z.B. durch Fragebögen) und vergleicht Gruppen hinsichtlich der Nutzung von adaptiven vs. maladaptiven Strategien.

ESM-Studien hingegen erfragen die Anwendung von Strategien über mehrere Tage oder Wochen via Smartphone in Situationen des täglichen Lebens. Daraus ergibt sich der einzigartige Vorteil, das dynamische Zusammenspiel von Situationen, Emotionen und Strategieanwendung in natürlichen Kontexten und frei von Erinnerungsverzerrungen zu erfassen, was die hohe ökologische Validität begründet, die ESM zugeschrieben wird. Beispielsweise werden Personen in «Laborstudien» dazu gebracht, Regulationsstrategien auszuwählen. Im realen Leben kann hingegen auch beobachtet werden, welche Situationen für eine Person überhaupt Emotions- oder Selbstregulation erfordern und inwiefern die gewählte Strategie für diese spezifische Situation angemessen ist (Benson et al., 2019; Blanke et al., 2020).

Die Anwendung von traditionellen Erhebungsmethoden führte zu einem statischen Konzept von Emotions- und Selbstregulationsstrategien, die oftmals eher pauschal als adaptiv oder maladaptiv eingeteilt wurden. Durch ESM und die Möglichkeit, Variabilität innerhalb von Personen und das Zusammenspiel mit variablen Kontexten zu analysieren hat sich in der Zwischenzeit eine neue, dynamische Perspektive eröffnet: Die Effektivität von Strategien wird jetzt im Zusammenhang mit der Passung zur Situation betrachtet. Personen mit einer guten Emotionsregulation haben eine hohe Variabilität (und somit ein breites Repertoire) zwischen verschiedenen Strategien, aber auch innerhalb von Strategien und sind in der Lage, die entsprechenden Strategien kontextsadäquat einzusetzen (Benson et al., 2019; Blanke et al., 2020; Wenzel et al., 2023). An diesem Beispiel wird deutlich, dass die Analyse von Variabilität innerhalb von Personen bedeutende Impulse für die Theoriebildung geben kann, sofern die entsprechenden Erhebungs- und Auswertungsmethoden zur Verfügung stehen.

Die sonderpädagogische Relevanz von IIV

Die beiden skizzierten Perspektivenwechsel haben Auswirkungen auf die sonderpädagogische Forschung und Praxis. Im Folgenden wird anhand der drei Bereiche Diagnostik, Förderung/Intervention und Forschung dargestellt, welche aktuellen Herausforderungen und zukünftigen Potentiale sich durch IIV spezifisch für sonderpädagogische Fachpersonen ergeben.

Diagnostik

Intraindividuelle Schwankungen sind ein inhärenter Bestandteil von menschlichem Verhalten und führen damit zu handfesten Problemen in der sonderpädagogischen Diagnostik, auf die Fachpersonen sensibilisiert werden sollten. Salthouse (2007) skizziert insbesondere zwei Schwierigkeiten: Zum einen kann eine einzelne statusdiagnostische Erhebung zu einer ungenauen Einschätzung der Leistung einer Person führen und im schlechtesten Fall sogar einen Extremwert repräsentieren, also die beste oder die schlechteste individuell mögliche Performanz anzeigen (siehe auch Ludwig et al., 2016). Möglicherweise kann damit sogar eine unterschiedliche diagnostische Klassifikation verbunden sein (z.B. eine zugesprochene oder nicht-zugesprochene Diagnose). Zum anderen wird die Interpretation von «within-person change», also von längerfristigen Entwicklungen, durch IIV in einzelnen Messungen erschwert. Was als (mangelnder oder vorhandener) Lernfortschritt interpretiert wird, kann auch auf kurzfristigen Leistungsfluktuationen beruhen. Da IIV bei bestimmten klinischen Gruppen erhöht ist, betreffen diese Probleme die Personen besonders, für die genaue diagnostische Informationen am wichtigsten sind. Wie oben bereits mehrfach erwähnt, zeigt sich IIV am offensichtlichsten dann, wenn dasselbe Verhalten über die Zeit mehrfach mittels gleichschwerer Paralleltests erfasst wird. In der sonderpädagogi-

schen Praxis werden solche Verfahren mit dem Curriculumbasierten Messen ([CBM], im deutschsprachigen Raum auch oft als Lernverlaufsdiagnostik (LVD) bezeichnet) auch eingesetzt. In der Literatur zu CBM/LVD wird IIV regelmässig thematisiert, weil sie bedeutsame Probleme für die Interpretation der Daten verursacht: Variabilität um den Fortschrittsgraphen führt dazu, dass mehr Messzeitpunkte benötigt werden, um einen stabilen Trend zu erkennen (Bundock et al., 2018; O’Keeffe et al., 2017). Dies wiederum kann ein Grund für inkorrekte pädagogische Entscheidungen über die Änderung oder die Fortführung der Instruktion sein, zumal Praktiker Probleme aufweisen, typischerweise auftretende Variabilität von Lernfortschritten zu unterscheiden (Toste et al., 2024) und auch über wenig Übung im Umgang mit Standardmessfehlern und Konfidenzintervallen verfügen (Fan & Hansmann, 2015). Verschiedene Lösungsansätze werden im Umgang mit messfehlerbedingter IIV in LVD/CBM-Daten empfohlen: Zum einen hat sich eine sorgfältige Testkonstruktion als notwendig erwiesen (O’Keeffe et al., 2017), die sich auch nicht ausschließlich auf die Paralleltestreliabilität verlassen darf, denn diese sichert nur die Vergleichbarkeit der Rangordnung der Schülerinnen und Schüler zwischen den verschiedenen Testformen ab, die gleichbleiben kann, wenn alle Kinder sich in dieselbe Richtung bewegen. Stattdessen können Verfahren wie die Generalisierbarkeitstheorie verwendet werden, welche einen absoluten Reliabilitätskoeffizienten verwenden und damit eine Vergleichbarkeit der Parallelformen auf dem individuellen Level gewährleisten (Fan & Hansmann, 2015). Die Reduktion von messfehlerbedingter IIV gelingt auch durch das Aggregieren von mehreren CBM-Tests, was erwiesenermaßen eine zuverlässigere Schätzung des «wahren» Wertes ermöglicht, aber zulasten der Ökonomie des Verfahrens geht (Ford & Kaldenberg, 2019). Ein Measurement-Burst-Design, wie es von Salthouse (2007) vorgeschlagen wird, könnte hier eine interessante Alternative darstel-

len: Dabei werden mehrere Assessments zur gleichen Zeit (also z.B. innerhalb eines oder weniger Tage) erhoben und diese Measurement-Bursts werden nach etwas längeren Zeiträumen (z.B. ein oder zwei Monate) wiederholt. Ob in diesem Zeitraum eine Veränderung stattgefunden hat, kann dann mittels individuell festgelegter t-Scores geprüft werden, was auch ohne Referenz zu Informationen von anderen Individuen möglich ist.

Förderung / Intervention

Mit dem Fokus auf IIV sind große Hoffnungen verbunden, Interventionen und Trainings adaptiv und individualisiert zu gestalten. Bei klassischen Interventionsstudien mit einem Prä-Posttestdesign stellt die Interventionsphase eine «Black Box» dar. Die Analyse von Variabilität innerhalb von Personen während des Trainings eröffnet hingegen Informationen über intraindividuelle Entwicklungsverläufe, Leistungsfluktuationen zwischen verschiedenen Trainingseinheiten und Verknüpfungen dieser Schwankungen mit anderen internen und externen Faktoren (Könen & Karbach, 2015). Welche Interventionen in welchen Kontexten für welche Personen wirksam sind, kann dadurch wesentlich exakter analysiert werden als durch eine Gruppenstudie mit lediglich zwei Messzeitpunkten (Beltz et al., 2016; Saqr, 2023, 2024). Könen und Karbach (2015) betonen, dass der Einbezug von within-Person Trainingsdaten ganz besonders geeignet ist für sehr heterogene Gruppen und für spezifische Populationen, wie z.B. Kinder mit Lernstörungen. Diese reagieren auf ein Training möglicherweise anders als Peers aus einer unauffälligen Kontrollgruppe, aber auch anders als andere Kinder mit Lernschwierigkeiten, abhängig von der Verursachung der Probleme und von vorher bereits erhaltenen Interventionen. Personalisierten Treatments, die auf idiographischen Analysen der Fluktuationen innerhalb von Individuen beruhen, wird aufgrund ihrer Genauigkeit und ihres hohen Neuigkeits-

gehalts ein großes Potential zugesprochen (Beltz et al., 2016; Saqr, 2024). Auch die Sonderpädagogik, innerhalb derer Individualisierung und Adaptivität traditionell einen hohen Stellenwert einnehmen, kann diese Informationsquelle nutzbar machen und somit neue Ansatzpunkte für gewinnbringende Förderung erschließen.

Eine praktische Anwendung von intensiven Datenerhebungsmethoden mit pädagogischen und therapeutischen Implikationen skizzieren Beltz et al. (2016). Sie zeigen anhand eines Beispiels von Personen mit psychischen Beeinträchtigungen, wie die tägliche Kovariation von internalisierenden und externalisierenden Verhaltensproblemen, erhoben mittels engmaschigen Verhaltensratings per Smartphone, für jeden einzelnen Patienten in einem Unified Structural Equation Model dargestellt werden kann. Solche Modelle beinhalten therapierelevantes Wissen, das eine individuell zugeschnittene Behandlung ermöglicht. ESM kann aber nicht nur als Erhebungsmethode dienen, sondern auch als Mittel der Intervention selbst in Form von Ecological Momentary Interventions (EMI), bei denen therapeutische Impulse oder Übungen via Smartphone im Alltag vermittelt werden und die optimalerweise adaptiv und individualisiert gestaltet sind. Bislang wird EMI vor allem in der Psychiatrie, Psychologie und für Interventionen im Bereich der mentalen Gesundheit diskutiert, die aber wiederum umfangreiche Überschneidungspunkte mit sonderpädagogischen Themen aufweisen (z.B. negativer Affekt, dysfunktionales Verhalten oder Social-Media-Use; Schick et al., 2023; Vries et al., 2021). Die Möglichkeiten, individualisierte Interventionen auf Basis der Analyse von intraindividuellen Daten zu erstellen, stecken noch in den Kinderschuhen und sind weit von einer niederschweligen Implementation in der Praxis entfernt. Hier eröffnet sich für die Sonderpädagogik ein aussichtsreiches Zukunftsfeld.

Forschung

Die Implikationen durch die beiden skizzierten Perspektivenwechsel sind umfangreich und werden die sonderpädagogische Forschung vermutlich nachhaltig beeinflussen. Im Folgenden werden einige mögliche Anknüpfungspunkte beispielhaft herausgegriffen. Die Erkenntnis, dass IIV Information darstellt, die analysiert werden kann und die beispielsweise als Prädiktor für Lernen und Entwicklung fungiert, führt zu einem vielschichtigen Bild: Je nach Aufgabenstellung wird eine erhöhte IIV entweder als Vulnerabilitätsfaktor und als Indikator für beeinträchtigte kognitive Funktionen interpretiert (z.B. bei Personengruppen mit sonderpädagogischem Förderbedarf, Lern- und Entwicklungsstörungen) oder aber als Anzeichen für fortschreitenden Strategie- und Kompetenzerwerb. Die spezifischen funktionalen Mechanismen zwischen IIV und Lernen und Verhalten sind noch nicht gut verstanden. So warnen Dirk und Schmiedek (2016) davor, aus korrelativen Resultaten von Gruppenstudien kausale Schlussfolgerungen zu ziehen und setzen sich dafür ein, IIV in theoretische Modelle von Lernen und Entwicklung zu integrieren. Ein möglicher Grund für das unvollständige Bild könnte darin bestehen, dass IIV in kognitiven Aufgabenstellungen zwar schon in vielen Studien mit unterschiedlichen Populationen untersucht wurde, es aber kaum Forschung zu IIV beispielsweise bei akademischen Leistungen selber gibt (Borella et al., 2011). Einen Grund dafür sehen Dirk und Schmiedek (2016) im Mangel an psychometrisch evaluierten Verfahren mit einer hohen Anzahl an vergleichbaren Testitems, weshalb Aufgabenstellungen zu kognitiven Grundfunktionen als Proxy für Schulleistungen genutzt werden. Die Sonderpädagogik könnte mit sorgfältig konstruierten LVD/CBM-Verfahren dazu beitragen, diesen wichtigen «missing link» zu schließen und das Bild durch weitere Facetten zu ergänzen. Bislang wird IIV in Lernverlaufsdaten mittels CBM-Verfahren noch nicht syste-

matisch als Informationsquelle betrachtet und analysiert. Dieses Desiderat erstaunt in Anbetracht der Tatsache, dass Fuchs et al. bereits im Jahr 1985 darauf hinwiesen, dass Performance Instability schon historisch gesehen als charakteristisches Merkmal von Kindern mit Learning Disabilities (LD) angesehen wurde (Fuchs et al., 1985). In zwei Studien (1985; 1986) untersuchte die Autorenschaft daher die performance instability von Schülerinnen und Schülern mit LD und von Schülerinnen und Schülern mit emotionalen Beeinträchtigungen beim Lesen mittels eines 1-Minute Lesetests. Zwischen beiden Gruppen gab es keine Unterschiede, was Fuchs et al. (1985) so interpretieren, dass IIV nicht als Signatur von LD fungiert und die Untersuchung von IIV keine praktische Relevanz aufweist. Aus heutiger Perspektive verwundert dieser Rückschluss aus Studien ohne unauffällige Kontrollgruppe nicht, da eine erhöhte IIV mit unterschiedlichen Entwicklungsauffälligkeiten assoziiert ist. Weitere Forschung zu dieser Thematik wurde im Anschluss daran dann aber lange Zeit nicht betrieben. In einer Studie zur IIV bei CBM-Schreiben (Winkes, eingereicht) demonstrierten schreibschwache Schülerinnen und Schüler eine höhere IIV als ihre schreibstarken Peers. Weiterhin erwies sich IIV von zehn Schreibproben im Herbst des Schuljahres in dieser Studie als Prädiktor für den relativen Leistungszuwachs in den nächsten sechs Monaten und erklärte dabei 23% der Varianz im Lernfortschritt. Diese ersten explorativen Ergebnisse sollten erweitert werden durch Studien mittels LVD/CBM in anderen akademischen Bereichen. Als besonders aufschlussreich könnte sich der Vergleich von IIV in Aufgaben herausstellen, die unterschiedliche Anforderungen stellen: Während die Leseflüssigkeit beispielsweise eine Fähigkeit darstellt, die weitestgehend automatisiert werden kann, ist die Schreibflüssigkeit auf die Anwendung unterschiedlicher Wissensquellen und Strategien angewiesen. Auch unterschiedliche Populationen könnten systematischer in die Forschung einbezogen werden, denn

die vorhandenen Kenntnisse zu IIV sind bei den meisten Lern- und Entwicklungsstörungen bis heute nur fragmentarisch. IIV besitzt das Potential, zum Wissen über Gemeinsamkeiten und Unterschiede zwischen verschiedenen Störungsbildern beizutragen und deren funktionale Mechanismen besser zu verstehen.

Der zweite Perspektivenwechsel erweitert die Möglichkeiten sonderpädagogischer Forschung fundamental. Die Chance, Smartphones für Datenerhebungen zu verwenden, hat in den vergangenen Jahren zu einer sehr regen Forschungstätigkeit mittels ESM-Verfahren geführt. Venetz und Zurbriggen (2015) berichten vor einem knappen Jahrzehnt, dass eine Literaturrecherche in den Datenbanken PsychInfo, ERIC und Medline etwa ein Dutzend sonderpädagogisch relevante empirische Studien erbrachte. Eine erneute Suche in den gleichen Datenbanken mit den Suchwörtern «Experience Sampling Method» und «Special Education» ergibt nun 324 Ergebnisse. ESM wird als ideale Linse bezeichnet, durch welche Variabilität betrachtet werden kann (Blanke et al., 2020). Für sonderpädagogisch relevante Populationen bringen sie weiterhin den Vorteil mit, dass sie kurz und einfach in der Handhabung sind und jeweils zeitnah zum interessierenden Geschehen ausgefüllt werden (Venetz & Zurbriggen, 2015). Sie können auch durch physiologische Daten (z.B. Herzrate, Schlafverhalten, sportliche Aktivitäten) ergänzt werden. Das Vorhandensein von Daten zu IIV führt allerdings nicht automatisch dazu, dass diese Wissensquelle auch ausgeschöpft wird. Viele ESM-Studien nutzen zwar Mehrebenenanalysen, fokussieren dabei aber immer noch auf Mittelwerte und lassen einen großen Anteil ihrer Daten ungenutzt (Schick et al., 2023; Vries et al., 2021). Dies dürfte auch mit der methodischen Komplexität der Daten in Zusammenhang stehen.

Fazit

Forschung und Praxis sehen sich mit bedeutsamen intra-individuellen Schwankungen in kognitiven Leistungen und im Verhalten von Personen konfrontiert. In verschiedenen kognitiven und neuropsychologischen Variablen beträgt die durchschnittliche intraindividuelle Standardabweichung ca. 50% der Standardabweichung zwischen Personen (Dirk & Schmiedek, 2016; Salt-house, 2007), was darauf hinweist, dass IIV ein Phänomen darstellt, welches nicht ignoriert werden kann und sollte. In anderen Disziplinen, wie der Psychologie, wächst die Einsicht, dass wir in einer dynamischen und komplexen Welt leben und dass daher statische Konzepte, z.B. von Intelligenz oder von Persönlichkeit vereinfachend sind. Es besteht deswegen auch oftmals eine schlechte Passung zwischen diesen Konzepten (und entsprechenden Testergebnissen) und dem Verhalten von Personen in der «echten Welt» (Vaughan & Birney, 2023). Beltz et al. (2016) legen am Beispiel von Personen mit psychischen Störungen dar, dass diese keine distinkten qualitativen Gruppen repräsentieren, die zeitlich stabil bestehen, sondern besser als Manifestationen von zeitlich variierenden Levels von multiplen, oft auch biologisch bedingten Verhaltensweisen beschreibbar sind. Auch auf die Sonderpädagogik ist diese Betrachtungsweise adaptierbar. Sie erfordert aber Anpassungen in den genutzten Erhebungs- und Analysemethoden. Der damit einhergehende Aufwand scheint lohnenswert, um neue Einsichten zu gewinnen, oder um es in den Worten von Boker et al. (2009, S. 858) auszudrücken: “The concerns are now less about *whether* variability within an individual should be studied than it is about *how* to make use of this important source of information about psychological processes.”

Literatur

- Ali, S., Kerns, K. A., Mulligan, B. P., Olson, H. C. & Astley, S. J. (2018). An investigation of intra-individual variability in children with fetal alcohol spectrum disorder (FASD). *Child Neuropsychology*, 24(5), 617–637. <https://doi.org/10.1080/09297049.2017.1302579>
- Amenta, S. & Crepaldi, D. (2016). Editorial: The Variable Mind? How Apparently Inconsistent Effects Might Inform Model Building. *Frontiers in psychology*, 7, 1–2. <https://doi.org/10.3389/fpsyg.2016.00185>
- Beck, E. D. & Jackson, J. J. (2021). Within-person variability. In J. F. Rauthmann (Hrsg.), *The handbook of personality dynamics and processes* (S. 75–100). Academic Press.
- Beltz, A. M., Wright, A. G. C., Sprague, B. N. & Molenaar, P. C. M. (2016). Bridging the Nomothetic and Idiographic Approaches to the Analysis of Clinical Data. *Assessment*, 23(4), 447–458. <https://doi.org/10.1177/1073191116648209>
- Benson, L., English, T., Conroy, D. E., Pincus, A. L., Gerstorff, D. & Ram, N. (2019). Age differences in emotion regulation strategy use, variability, and flexibility: An experience sampling approach. *Developmental psychology*, 55(9), 1951–1964. <https://doi.org/10.1037/dev0000727>
- Blanke, E. S., Brose, A., Kalokerinos, E. K., Erbas, Y., Riediger, M. & Kuppens, P. (2020). Mix it to fix it: Emotion regulation variability in daily life. *Emotion*, 20(3), 473–485. <https://doi.org/10.1037/emo0000566>
- Blume, F., Irmer, A., Dirk, J. & Schmiedek, F. (2022). Day-to-day variation in students' academic success: The role of self-regulation, working memory, and achievement goals. *Developmental science*, 25(6), e13301. <https://doi.org/10.1111/desc.13301>
- Boker, S. M., Molenaar, P. C. M. & Nesselroade, J. R. (2009). Issues in intraindividual variability: individual differences in equilibria and dynamics over multiple time scales. *Psychology and aging*, 24(4), 858–862. <https://doi.org/10.1037/a0017912>
- Borella, E., Chicherio, C., Re, A. M., Sensini, V. & Cornoldi, C. (2011). Increased intraindividual variability is a marker of ADHD but also of dyslexia: A study on handwriting. *Brain and Cognition*, (77), 33–39.
- Bulté, B. & Housen, A. (2020). A critical appraisal of the CDST approach to investigating linguistic complexity in L2 writing development. In G. G. Fogal & M. H. Verspoor (Hrsg.), *Language Learning & Language Teaching. Complex Dynamic Systems Theory and L2 Writing Development* (Bd. 54, S. 207–238). John Benjamins Publishing Company. <https://doi.org/10.1075/llt.54.09bul>
- Bundock, K., O'Keeffe, B. V., Stokes, K. & Kladis, K. L. (2018). Strategies for Minimizing Variability in Progress Monitoring of Oral Reading Fluency. *TEACHING Exceptional Children*, 50(5), 273–281. <https://doi.org/10.1177/0040059918764097>
- Burton, C. L., Strauss, E., Hultsch, D. F. & Hunter, M. A. (2009). The relationship between everyday problem solving and inconsistency in reaction time in older adults. *Aging, Neuropsychology and Cognition*, 16(5), 607–632. <https://doi.org/10.1080/13825580903167283>
- Cañigueral, R., Ganesan, K., Smid, C. R., Thompson, A., Dosenbach, N. U. F. & Steinbeis, N. (2023). Intra-individual variability adaptively increases following inhibition training during middle childhood. *Cognition*, 239, 105548. <https://doi.org/10.1016/j.cognition.2023.105548>
- Castellanos, F. X., Sonuga-Barke, E. J. S., Scheres, A., Di Martino, A., Hyde, C. & Walters, J. R. (2005). Varieties of attention-deficit/hyperactivity disorder-related intra-individual variability. *Biological psychiatry*, 57(11), 1416–1423. <https://doi.org/10.1016/j.biopsych.2004.12.005>
- Cubillo, A., Hermes, H., Berger, E., Winkel, K., Schunk, D., Fehr, E. & Hare, T. A. (2023). Intra-individual variability in task performance after cognitive training is associated with long-term outcomes in children. *Developmental Science*, 26(1), e13252. <https://doi.org/10.1111/desc.13252>
- Diehl, M., Hooker, K. & Sliwinski, M. J. (2015). A Brief Historical Overview of Intraindividual Variability Research Across the Life Span. In M. Diehl, K. Hooker & M. J. Sliwinski (Hrsg.), *Handbook of Intraindividual Variability Across the Life Span* (S. 3–15). Routledge.
- Dirk, J. & Schmiedek, F. (2016). Fluctuations in elementary school children's working memory performance in the school context. *Journal of Educational Psychology*, 108(5), 722–739. <https://doi.org/10.1037/edu0000076>

- Dirk, J. & Schmiedek, F. (2017). Variability in Children's Working Memory Is Coupled With Perceived Disturbance: An Ambulatory Assessment Study in the School and Out-of-School Context. *Research in Human Development, 14*(3), 200–218. <https://doi.org/10.1080/15427609.2017.1340051>
- Fagot, D., Mella, N., Borella, E., Ghisletta, P., Lecerf, T. & Ribaupierre, A. de (2018). Intra-Individual Variability from a Lifespan Perspective: A Comparison of Latency and Accuracy Measures. *Journal of Intelligence, 6*(1). <https://doi.org/10.3390/jintelligence6010016>
- Fan, C.-H. & Hansmann, P. R. (2015). Applying Generalizability Theory for Making Quantitative RTI Progress-Monitoring Decisions. *Assessment for Effective Intervention, 40*(4), 205–215. <https://doi.org/10.1177/1534508415573299>
- Ford, J. W. & Kaldenberg, E. R. (2019). Curriculum-Based Measurement for Written Expression with Postsecondary Students with Intellectual and Developmental Disabilities. *Journal of Inclusive Postsecondary Education, 1*(2), 1–22. <https://doi.org/10.13021/JIPE.2019.2473>
- Fuchs, D., Fuchs, L. S. & Deno, S. L. (1985). Performance Instability: An Identifying Characteristic of Learning Disabled Children? *Learning Disability Quarterly, 8*(1), 19–26.
- Fuchs, D., Fuchs, L. S., Tindal, G. & Deno, S. L. (1986). Performance Instability of Learning Disabled, Emotionally Handicapped, and Nonhandicapped Children. *Learning Disability Quarterly, 9*(1), 84–88.
- Galeano-Keiner, E. M., Neubauer, A. B., Irmer, A. & Schmiedek, F. (2022). Daily fluctuations in children's working memory accuracy and precision: Variability at multiple time scales and links to daily sleep behavior and fluid intelligence. *Cognitive Development, 64*, 101260. <https://doi.org/10.1016/j.cogdev.2022.101260>
- Geurts, H. M., Grasman, R. P. P., Verté, S., Oosterlaan, J., Roeyers, H., van Kammen, S. M. & Sergeant, J. A. (2008). Intra-individual variability in ADHD, autism spectrum disorders and Tourette's syndrome. *Neuropsychologia, 46*(13), 3030–3041. <https://doi.org/10.1016/j.neuropsychologia.2008.06.013>
- Gooch, D., Snowling, M. J. & Hulme, C. (2012). Reaction time variability in children with ADHD symptoms and/or dyslexia. *Developmental neuropsychology, 37*(5), 453–472. <https://doi.org/10.1080/87565641.2011.650809>
- Huang, T., Steinkrauss, R. & Verspoor, M. (2021). Variability as predictor in L2 writing proficiency. *Journal of Second Language Writing, 52*, 100787. <https://doi.org/10.1016/j.jslw.2020.100787>
- Isbell, E., Calkins, S. D., Swingler, M. M. & Leerkes, E. M. (2018). Attentional fluctuations in preschoolers: Direct and indirect relations with task accuracy, academic readiness, and school performance. *Journal of experimental child psychology, 167*, 388–403. <https://doi.org/10.1016/j.jecp.2017.11.013>
- Judd, N., Klingberg, T. & Sjöwall, D. (2021). Working memory capacity, variability, and response to intervention at age 6 and its association to inattention and mathematics age 9. *Cognitive Development, 58*, 1–9. <https://doi.org/10.1016/j.cogdev.2021.101013>
- Kautto, A., Railo, H. & Mainela-Arnold, E. (2024). Introducing the Intra-Individual Variability Hypothesis in Explaining Individual Differences in Language Development. *Journal of Speech, Language, and Hearing Research, 67*(8), 2698–2707. https://doi.org/10.1044/2024_JSLHR-23-00527
- Könen, T. & Karbach, J. (2015). The benefits of looking at intraindividual dynamics in cognitive training data. *Frontiers in psychology, 6*, 615. <https://doi.org/10.3389/fpsyg.2015.00615>
- Lievens, F., Lang, J. W. B., Fruyt, F. de, Corstjens, J., van de Vijver, M. & Bledow, R. (2018). The predictive power of people's intraindividual variability across situations: Implementing whole trait theory in assessment. *The Journal of Applied Psychology, 103*(7), 753–771. <https://doi.org/10.1037/apl0000280>
- Lowie, W. M. & Verspoor, M. H. (2019). Individual Differences and the Ergodicity Problem. *Language Learning, 69*, 184–206. <https://doi.org/10.1111/lang.12324>
- Ludwig, K., Haindl, A., Laufs, R. & Rauch, W. A. (2016). Self-Regulation in Preschool Children's Everyday Life: Exploring Day-to-Day Variability and the Within- and Between-Person Structure. <https://doi.org/10.11588/josar.2016.2.34357>
- Molenaar, P. C. M. (2004). A Manifesto on Psychology as Idiographic Science: Bringing the Person Back Into Scientific Psychology, This Time Forever.

- Measurement: Interdisciplinary Research & Perspective*, 2(4), 201–218. https://doi.org/10.1207/s15366359mea0204_1
- Molenaar, P. C. & Campbell, C. G. (2009). The New Person-Specific Paradigm in Psychology. *Current Directions in Psychological Science*, 18(2), 112–117. <https://doi.org/10.1111/j.1467-8721.2009.01619.x>
- Nesselroade, J. R. & Molenaar, P. C. M. (2010). Emphasizing Intraindividual Variability in the Study of Development Over the Life Span. In R. M. Lerner, M. E. Lamb & A. M. Freund (Hrsg.), *The Handbook of Life-Span Development* (Bd. 28, S. 1). John Wiley & Sons <https://doi.org/10.1002/9780470880166.hlsd001002>
- O’Keeffe, B. V., Bundock, K., Kladis, K. L., Yan, R. & Nelson, K. (2017). Variability in DIBELS Next Progress Monitoring Measures for Students at Risk for Reading Difficulties. *Remedial and Special Education*, 38(5), 272–283. <https://doi.org/10.1177/0741932517713310>
- Salthouse, T. A. (2007). Implications of within-person variability in cognitive and neuropsychological functioning for the interpretation of change. *Neuropsychology*, 21(4), 401–411. <https://doi.org/10.1037/0894-4105.21.4.401>
- Saqr, M. (2023). Modelling within-person idiographic variance could help explain and individualize learning. *British Journal of Educational Technology*, 54(5), 1077–1094. <https://doi.org/10.1111/bjet.13309>
- Saqr, M. (2024). Group-level analysis of engagement poorly reflects individual students’ processes: Why we need idiographic learning analytics. *Computers in Human Behavior*, 150, 107991. <https://doi.org/10.1016/j.chb.2023.107991>
- Schick, A., Rauschenberg, C., Ader, L., Daemen, M., Wieland, L. M., Paetzold, I., Postma, M. R., Schulte-Strathaus, J. C. C. & Reininghaus, U. (2023). Novel digital methods for gathering intensive time series data in mental health research: scoping review of a rapidly evolving field. *Psychological Medicine*, 53(1), 55–65. <https://doi.org/10.1017/S0033291722003336>
- Schmiedek, F., Lövdén, M. & Lindenberger, U. (2009). On the relation of mean reaction time and intraindividual reaction time variability. *Psychology and Aging*, 24(4), 841–857. <https://doi.org/10.1037/a0017799>
- Tamm, L., Epstein, J. N. & Becker, S. P. (2019). A preliminary investigation of reaction time variability in relation to social functioning in children evaluated for ADHD. *Child Neuropsychology*, 25(7), 885–898. <https://doi.org/10.1080/09297049.2018.1523379>
- Toste, J. R., Filderman, M. J., Clemens, N. H. & Fry, E. (2024). Graph Out Loud: Pre-Service Teachers’ Data Decisions and Interpretations of CBM Progress Graphs. *Journal of Learning Disabilities*. <https://doi.org/10.1177/00222194241231768>
- van Dijk, M., Lowie, W., Smit, N., Verspoor, M. & van Geert, P. [Paul] (2024). Complex dynamic systems theory as a foundation for process-oriented research on second language development. *Second Language Research*, <https://doi.org/10.1177/02676583241246739>
- van Dijk, M. & van Geert, P. (2015). The Nature and Meaning of Intraindividual Variability in Development in the Early Life Span. In M. Diehl, K. Hooker & M. J. Sliwinski (Hrsg.), *Handbook of intraindividual variability across the life-span* (S. 37–58). Routledge Taylor & Francis Group.
- van Geert, P. & van Dijk, M. (2002). Focus on variability: New tools to study intra-individual variability in developmental data. *Infant Behavior and Development*, 25(4), 340–374. [https://doi.org/10.1016/S0163-6383\(02\)00140-6](https://doi.org/10.1016/S0163-6383(02)00140-6)
- Vaughan, A. C. & Birney, D. P. (2023). Within-Individual Variation in Cognitive Performance Is Not Noise: Why and How Cognitive Assessments Should Examine Within-Person Performance. *Journal of Intelligence*, 11(6). <https://doi.org/10.3390/jintelligence11060110>
- Venetz, M. & Zurbruggen, C. (2015). Intensive Longitudinal Methods – ihre Eignung für die sonderpädagogische Forschung und exemplarische Anwendungsmöglichkeiten. *Empirische Sonderpädagogik*, 7, 194–205. <https://doi.org/10.25656/01:11381>
- Verspoor, M. & de Bot, K. (2021). Measures of variability in transitional phases in second language development. *International Review of Applied Linguistics in Language Teaching*. <https://doi.org/10.1515/iral-2021-0026>
- Vries, L. P. de, Baselmans, B. M. L. & Bartels, M. (2021). Smartphone-Based Ecological Momentary Assessment of Well-Being: A Systematic Review and Recommendations for Future Studies.

Journal of Happiness Studies, 22(5), 2361–2408. <https://doi.org/10.1007/s10902-020-00324-7>

Wang, Z., Sun, Y., Wang, H., Liu, Q. & Zhang, R. (2022). Intraindividual reaction time variability and academic achievement of primary school students: A longitudinal cross-lagged panel model. *Journal of Applied Developmental Psychology*, 82, 101454. <https://doi.org/10.1016/j.appdev.2022.101454>

Wenzel, M., Bürgler, S., Rowland, Z. & Hennecke, M. (2023). Self-Control Dynamics in Daily Life: The Importance of Variability Between Self-Regulatory Strategies and Strategy Differentiation. *European Journal of Personality*, 37(1), 33–56. <https://doi.org/10.1177/08902070211043023>

Wolfe, K. R., Walsh, K. S., Reynolds, N. C., Mitchell, F., Reddy, A. T., Paltin, I. & Madan-Swain, A. (2013). Executive functions and social skills in survivors of pediatric brain tumor. *Child neuropsychology : a journal on normal and abnormal development in childhood and adolescence*, 19(4), 370–384. <https://doi.org/10.1080/09297049.2012.669470>

Zeng, S., Wang, Z., Zhang, R. & Yang, X. (2022). The Longitudinal Relationships between Family Functioning and Children’s Conduct Problems: The Moderating Role of Attentional Control Indexed

by Intraindividual Response Time Variability. *Journal of Psychopathology and Behavioral Assessment*, 44(3), 862–874. <https://doi.org/10.1007/s10862-022-09964-w>

Autorinnen- und Autorenhinweis

 Julia Winkes
<https://orcid.org/0000-0002-4383-1926>

Korrespondenzadresse

Julia Winkes
Universität Freiburg
Petrus-Kanisius-Gasse 21
CH-1700 Freiburg
julia.winkes@unifr.ch

Erstmals eingereicht: 28.10.2024
Überarbeitung eingereicht: 27.04.2025
Angenommen: 26.05.2025

Offene Daten	Es wurden keine Daten verwendet.
Offener Code	Es wurde kein Computercode verwendet.
Offene Materialien	NA
Präregistrierung	Nein
Votum Ethikkommission	Nein
Finanzielle und weitere sachliche Unterstützung	Es erfolgte keine finanzielle und weitere sachliche Unterstützungen durch Dritte.
Autorenschaft	JW hat den gesamten Beitrag alleine geschrieben.
KI und generative Modelle	Chat-GPT wurde verwendet, um Ideen für den Titel zu generieren und für die Umformulierung einzelner Sätze. Chat-GPT und DeepL wurden für die Übersetzung des Abstracts auf Englisch genutzt.

Intra-individual Differences in Situational Appraisals of Mathematics Anxiety in Grade Four Students

Moritz Herzog¹, Miriam Balt¹ & Carmen Zurbriggen²

¹ University of Cologne

² University of Fribourg

Abstract

Mathematics anxiety (MA) affects a significant number of students in primary schools. Recent research has emphasized the importance of situational and subjective assessment of MA. In particular, the relevance of a mathematical situation as well as the difficulty of tasks affect students' experience of MA. Whereas a growing body of studies examine inter-individual variance in MA experience, research on intra-individual MA is still scarce. The experience sampling method (ESM) appears to be a suitable approach to address intra-individual differences in MA. The current study aims at investigating associations of situational characteristics and MA experience with particular focus on gender differences. A total of 57 students from four Grade 4 classes were followed over the course of one lesson. Students' MA, their subjective evaluation of situation relevance, and task difficulty were assessed with ESM four times during each lesson at semi-randomized time points. Data was analyzed with multilevel regression models. Results showed that higher task difficulty was associated with an increased MA. Situational relevance showed a gender-specific association with MA: Girls reported significantly higher MA in more relevant situations, while higher MA in boys was (not significantly) associated with less relevant situations. The results bolster the importance of in situ assessment of MA focusing on intra-individual variance. Gender differences in the situational interplay suggest that task difficulty is evaluated by all students based on similar information, whereas situational relevance is evaluated differently by boys and girls.

Keywords: mathematics anxiety, gender differences, mathematics education, experience sampling methods, multilevel analysis

Intra-individuelle Unterschiede in der situativen Bewertung von Mathematikangst bei Lernenden der vierten Klasse

Zusammenfassung

Mathematikangst (MA) betrifft eine beträchtliche Anzahl von Schüler*innen bereits im Grundschulalter. Aktuelle Forschungsergebnisse haben die Bedeutung einer situativen und subjektiven Bewertung von MA hervorgehoben. Insbesondere die Relevanz der mathema-

tischen Situation sowie der Schwierigkeitsgrad der Aufgaben beeinflussen die MA-Erfahrung der Schüler*innen. Während sich zahlreiche Studien mit der interindividuellen Varianz des Erlebens von MA befassen, gibt es derzeit nur wenige Forschungsarbeiten zur intraindividuellen MA. Experience Sampling Methods (ESM) scheinen ein geeigneter Ansatz zu sein, um intraindividuelle Unterschiede in der MA zu untersuchen. Die vorliegende Studie zielt darauf ab, Zusammenhänge zwischen situativen Merkmalen und MA-Erfahrungen zu untersuchen, wobei ein besonderer Schwerpunkt auf geschlechtsspezifischen Unterschieden liegt. Insgesamt nahmen 57 Schüler*innen aus vier Klassen der vierten Jahrgangsstufe teil. Während jeder Unterrichtsstunde wurden das Erleben von MA der Schüler*innen, ihre subjektive Bewertung der Relevanz der Situation und der Schwierigkeitsgrad der Aufgabe viermal zu teilrandomisierten Zeitpunkten im Selbstbericht erfasst. Die Daten wurden mit Mehrebenen-Regressionsmodellen analysiert. Die Ergebnisse zeigen, dass ein höherer Schwierigkeitsgrad der Aufgabe mit erhöhter MA verbunden war. Die Relevanz der Situation zeigte einen geschlechtsspezifischen Zusammenhang mit MA: Mädchen berichteten über signifikant größere MA in relevanteren Situationen, während größere MA bei Jungen (nicht signifikant) mit weniger relevanten Situationen verbunden war. Die Ergebnisse unterstreichen die Bedeutung einer In-situ-Bewertung von MA mit Fokus auf intraindividuelle Varianz. Geschlechtsspezifische Unterschiede im situativen Zusammenspiel deuten darauf hin, dass die Schwierigkeit der Aufgabe von allen Schüler*innen auf der Grundlage ähnlicher Informationen bewertet wird, während die Situationsrelevanz von Jungen und Mädchen unterschiedlich bewertet wird.

Schlagerwörter: Mathematikangst, Geschlechtsunterschiede, Mathematiklernen, experience sampling methods, Mehrebenenanalysen

The issue of mathematics anxiety (MA) in primary school students has recently become a topic of growing interest among researchers, and particular interest has been devoted to the assessment of MA (e.g., Orbach et al., 2020; 2025; Ramirez et al., 2018). In numerous studies, MA is assessed through the administration of a questionnaire wherein students are instructed to imagine themselves in a mathematical situation (e.g., completing mathematics homework) and then asked to indicate how they would feel in that situation (Orbach et al., 2019a). Retrospective responses provided by students regarding their feelings can be influenced by their attitudes and beliefs, so-called “recall bias” (Schmidt et al., 2023). Studies have provided evidence for a recall bias in MA assessments (O’Leary et al., 2017). Orbach et al. (2019a) advocate for the assessment of MA during a mathematical situation (e.g., when stu-

dents are engaged in homework) and the solicitation of students’ immediate feelings to obtain a more accurate representation of their actual experience of MA. However, situation-based assessments of MA do not account for differences in individual experience of MA between different situations over time. Is the homework a significant challenge or is it easy to complete? Will my the homework affect my mathematics grade, or will my teacher not even check?

The individual and subjective experience of MA may vary depending on the relevance and difficulty of the tasks in question (Ramirez et al., 2018). Consequently, the assessment of MA in a single situation provides only limited information regarding the individual emotions and attitudes towards mathematics of a student. Accordingly, it is essential to conduct multiple *in situ* assessments of MA for each student in order to account for intra-individual variance, too.

One effective methodological approach for achieving this is the use of the Experience Sampling Method (ESM; Hektner et al., 2007). ESM is designed to capture people's subjective, and situational experiences in their real-life context over an extended period of time (e.g., several days or weeks).

The present study seeks to contribute to the existing literature by investigating both intra- and inter-individual variance in students' experience of state-MA. To the best of our knowledge, the present study is the first to employ an ESM approach to MA, focusing on the individual impact of situation relevance and task difficulty on state-MA in varying mathematics classroom situations. It also addresses inter-individual variance in the form of potential gender differences in situational appraisals (Ramirez et al., 2018).

Mathematics anxiety

MA is frequently defined as „feelings of apprehension and tension concerning manipulation of numbers and completion of mathematical problems“ (Richardson & Suinn, 1972, p. 551), which affects individuals in academic and everyday contexts. In the case of students, mathematical situations lay mostly within the context of mathematics classes. Consequently, the investigation of MA in mathematics classrooms represents a significant area of interest and a primary focus of this study.

MA is associated with the typical symptoms of anxiety, including mental (e.g., worry, sadness) and physical stress indicators (e.g., increased heart rate, sweating, pain) that are experienced in anticipation or during mathematical situations. Furthermore, MA affects behavior, with students experiencing it displaying a tendency to avoid mathematical situations (Haase et al., 2019). While MA meets the criteria for a specific phobia, it has not yet been classified as a specific anxiety disorder (Orbach et al., 2019a). From a clinical perspective, math anxiety can be recognized as a dis-

tinct condition rather than a subset of general or test anxiety (Orbach et al., 2019a). This distinction is supported by research showing that correlations between various measures of math anxiety are stronger than those between math anxiety and other anxiety disorders, such as general or test anxiety (Hembree, 1990; Hopko, Mahadevan, Bare, & Hunt, 2003; Schnell et al., 2013). Additionally, Kazelskis (2000) reinforced the conceptual uniqueness of the construct math anxiety.

State- and trait-mathematics anxiety

An important aspect of anxiety is the differentiation of anxiety as a current feeling (state) and as a personal characteristic (trait; Spielberger, 1972). In the context of MA, state-MA refers to the anxiety experienced in mathematical situations, whereas trait-MA is operationalized by the fear of failure in mathematics (Sorvo et al., 2017). In line with the classic state-trait model of anxiety (Spielberger, 1972), Orbach et al. (2019b) put forth a state-trait model of MA (Fig. 1). State-MA is defined as the experience of mental stress and physical arousal during a mathematical situation (e.g., in mathematics classes). In contrast, trait-MA represents the general disposition and self-image of students when imagining mathematical situations. State-MA and trait-MA are conceptually interrelated. Those students who experience state-MA on a regular basis internalize the fears they have experienced and develop a self-image of being mathematics anxious. Consequently, they are more sensitive to mathematical situations and perceive them as a threat more frequently (Orbach et al., 2019b). The interplay of state- and trait-MA can form a vicious circle of self-perpetuating increase (Zeidner & Matthews, 2011).

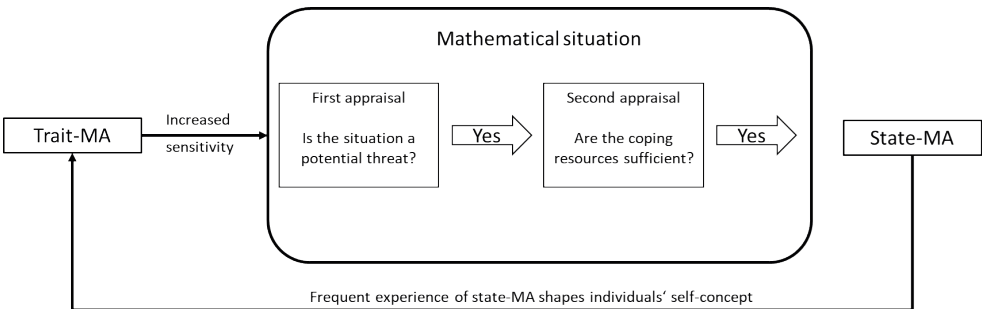


Figure 1
The state-trait model of mathematics anxiety (adopted from Orbach et al., 2019)

The development of MA and its relation-
ship to math performance

To explain the development of MA and its link to math performance, various theoretical perspectives have been proposed. The *reduced competency account* posits that individuals with lower mathematical abilities are more prone to experience MA, suggesting that lower mathematics competencies lead to MA (Ramirez et al, 2018). In addition, mathematics-anxious students tend to avoid mathematical situations, resulting in a lack of learning opportunities. This creates a self-reinforcing cycle where discouraging experiences in mathematics and avoidance behavior strengthen each other (Dowker et al., 2016). In this context, the *disruption account* argues that MA impairs executive functions and working memory resources (Moran, 2016). Rather than focusing on the task at hand, individuals with MA direct their attention to their emotions and affect during mathematical situations.

Despite a growing body of research on the relationship between MA and mathematical performance, findings remain inconclusive regarding a consistent negative association between the two (e.g., Dowker et al., 2016; Orbach et al., 2020). While meta-analyses report a small but significant negative correlation (Barroso et al., 2021; Nakung et al., 2019), other studies have failed to confirm this pattern (Haase et al., 2012; Krinzing-er et al., 2007; Thomas & Dowker, 2000; Wood et al., 2012).

The role of situational appraisals

The *cognitive appraisal theory* (Lazarus, 2001) explains the occurrence of MA based on two main cognitive appraisals (Fig. 1). Upon encountering a mathematical situation, individuals (mostly unconsciously) evaluate two questions regarding the situation. The first appraisal entails an evaluation of the situation's relevance. This initial cognitive appraisal determines whether the mathematical task is perceived as a potential threat, a challenge, or irrelevant. In the event that the situation is deemed irrelevant or a mere challenge, the probability of the individual experiencing MA is low. Conversely, if the situation is perceived as a potential threat, the second appraisal is evaluated. In this second appraisal, individuals compare the difficulty of the task with their individual mathematics skills. The less individuals perceive themselves as having the necessary skills to handle the task, the more likely they are to experience anxiety, and the more intense it becomes in that situation (Orbach et al., 2019b). There are various sources of relevance, independent of actual mathematical abilities, that can lead to evaluating a mathematical situation as a potential threat, such as the teacher's evaluation, the impression made on peers, or self-image. For example, a potential negative evaluation by the teacher could represent a threat, while the fear of embarrassment in front of peers might not (or vice versa, or both). The onset of anxiety in

a mathematical situation can therefore vary not only between students but also within the same student.

The *interpretation account*, proposed by Ramirez et al. (2018), expands upon cognitive appraisal theory by highlighting that the way students interpret their experiences in mathematical situations plays a crucial role in determining the extent to which they experience MA. This can include factors such as attributing past math performance to (lack of) ability or luck, feedback received from teachers, or physical reactions (e.g., increased heart rate) when solving math tasks. Therefore, the interpretation account can explain intra-individual fluctuations in appraisals across different situations (Ramirez et al., 2018). For example, individuals who constantly attribute poor performance to a lack of ability may be more prone to developing trait-MA, leading to lower intra-individual variability, as nearly every mathematical situation may be perceived as a threat.

The role of gender

The question of gender differences in MA has been explored in numerous studies, but the results are inconsistent. Theoretically, it is often suggested that social stereotypes, such as the belief that women are less talented in mathematics, may influence the perceptions and behaviors of girls, leading them to internalize these stereotypes (Beilock et al., 2012). These stereotypes might result in higher levels of MA in girls. Additionally, the deficit competency model, which posits that girls tend to perform worse in mathematics, could also contribute to increased MA among girls (Maloney et al., 2012).

Empirical studies, however, provide mixed evidence on this issue. Some studies have found higher levels of MA in girls, particularly in primary school (e.g., Carey et al., 2017; Hill et al., 2016; Krinzinger et al., 2007), while others report no significant gender differences (e.g., Dowker

et al., 2012; Wu et al., 2014). These discrepancies may arise from different ways in which MA is measured, especially the distinction between trait-MA (a stable anxiety) and state-MA (anxiety experienced in specific situations). For example, studies that identify gender differences in MA often focus on trait-MA and fear of failure (Hill et al., 2016; Krinzinger et al., 2007), whereas when MA is measured during mathematical tasks (state-MA), gender seems to have little to no impact on anxiety levels (e.g., Wu et al., 2014).

In this context, Ashcraft's (2002) findings provide further insight. He observed that female participants were more likely to report anxiety in general, which suggests that the apparent gender differences in MA might not reflect an inherent disparity in anxiety levels between the genders. Rather, it may indicate that girls are more likely to express their anxiety openly. This idea aligns with the observed discrepancy in findings between trait- and state-MA measures, where the gender gap in reported anxiety becomes less pronounced when anxiety is assessed in specific performance situations (Orbach et al., 2019a).

Assessing MA with the ESM

A review of extant research reveals the existence of at least two notable distinctions in the manner in which MA is assessed. First, studies that assess fear of failure report a weaker association between anxiety and performance than those that assess anxiety about mathematical situations (Sorvo et al., 2017). Second, MA appears to be more strongly associated with lower mathematics performance when it is assessed during mathematical situations (Orbach et al., 2019b; 2020). Consequently, Orbach et al. (2020) posit that the disparate findings regarding the link between MA and mathematics performance can be at least partially explained by differences in assessment methods (Dowker et al., 2016).

The ESM aims to assess people's subjective, momentary experience in real time and in their real-life context (i.e., in situ), which allows to capture state-like constructs such as positive activation or MA and to link them with characteristics of the corresponding situational context (Hektner et al., 2007). The main idea of this assessment approach is to draw a representative sample of occasions or 'snapshots' of a specific aspect of a person's everyday life according to the respective research question. This is achieved by assessing not only in real-time and in situ but also several times over a certain relatively short period of time (i.e., several days or weeks), which further enables to examine intraindividual variability and the fluctuating nature of state-like constructs. Given the repeated measurement design, the ESM belongs to the group of intensive longitudinal methods (Bolger & Laurenceau, 2013).

Two of the main advantages of the ESM are the high ecological validity and lower retrospection effects compared to the retrospective assessment of conventional questionnaires (e.g., Venetz & Zurbriggen, 2015). Previous studies using the ESM during lessons have indicated that students' retrospective reports of their emotional experience in the classroom are affected by relatively stable characteristics such as students' attitudes toward school (Zurbriggen et al., 2021) and personality traits (Schmidt et al., 2023). Two disadvantages of the ESM are the increased effort for participants and the complexity of the data structure.

The present study

The literature review highlights the need for further exploration of the situational aspects of MA across multiple levels, including the model level (i.e., the state-trait model), the performance level (i.e., the anxiety-performance link in state-like MA assessments), and the gender level (i.e., gender differences in state-like MA assessments). MA involves both inter- and intra-individual variance,

as different students may react differently to the same situation, and the same student may respond to different situations in varying ways (Ramirez et al., 2018). Given the inconsistencies in MA research, future studies should more thoroughly investigate these individual differences.

The present study aligns with the state-trait model and the interpretation account, focusing on intra-individual variance in students' experiences across mathematical situations. Since findings on gender differences remain inconsistent and recent studies suggest that in-situ assessments provide clearer insights (Orbach et al., 2019a), our aim is to explore gender differences in the interaction between situation relevance, task difficulty, and their impact on MA.

In this study *situation relevance* refers to how significant or important a situation is perceived by an individual in relation to his or her personal goals, values, or well-being (Prinski et al., 2018). It is assumed that a situation perceived as relevant is associated with a certain effort required to handle it (Ainley, 2012). Additionally, relevant situations are often situations, in which making no mistakes is crucial (e.g., exams). *Task difficulty* describes the perceived level of difficulty a situation presents in relation to the individual's abilities. A task is considered difficult if it exceeds one's perceived capacity to handle it successfully. Note that the individually perceived difficulty of a task may not necessarily be in line with estimated difficulties based on empirical data (e.g., in IRT models), but is grounded by the students' individual perspective (Ramirez et al., 2018).

The following two research questions (RQ) guide the current study:

(RQ1) Are situational characteristics, particularly the relevance of the situation and the difficulty of the task, associated with the students' experience of MA in mathematics classrooms? Given natural variance in situational relevance and task difficulty during classes, the intra-individual experience is of great interest (Ramirez et al., 2018).

Based on the state-trait model of mathematics anxiety (Orbach et al., 2019b), we hypothesize that students experience MA more frequently and intensely when (a) the situation is more relevant and (b) the task is more difficult. More relevant situations are more likely to trigger the first appraisal which initiates a potential MA response. More difficult tasks are more likely to exceed students' mathematics skills, affecting the second appraisal (Orbach et al., 2019b). As the appraisals are described as distinct processes, we thrive to analyze the individual association of situation relevance and task difficulty on state-MA separately, but not the concurrent association of both.

(RQ2) In which way do girls and boys differ in terms of the association between situational characteristics (i.e., relevance and difficulty of a task) and their experience of MA? Empirical evidence does not provide a consistent picture regarding gender differences in state-MA. The situational interplay of relevance and task difficulty in boys and girls might provide further insights in this regard. Some researchers hypothesize that social factors lead to higher MA in girls (e.g., Ashcraft, 2002; Beilock et al., 2012), while others propose that gender differences are lower in state-MA assessments compared to trait-MA (Orbach et al., 2019b). Given the contradictory findings regarding gender differences in empirical studies on state- and trait-MA (e.g., Hill et al., 2016; Wu et al., 2014) and the potentially complex interactions of gender, situation relevance, task difficulty, and MA, we refrain from stating a priori hypotheses regarding gender differences in the effects of situational characteristics on state-MA.

Materials and Methods

Design and procedure

The study employed the ESM approach to investigate MA in the context of mathematics classrooms. Four classes were surveyed

over the course of a series of lessons on a specific topic. The lesson series spanned a duration of between nine and twelve lessons. The duration of each lesson was 45 minutes. In order to encompass a comprehensive range of potential scenarios within the context of mathematics classes, students' situational experiences were evaluated on four occasions per lesson throughout the course of the lesson series.

The signals were distributed semi-randomly across the duration of a single lesson (i.e., semi-random sampling scheme; Dejonckheere & Erbas, 2021). The initial five minutes were excluded from the analysis, as teachers typically engage in organizational and preparatory tasks during this period. It seems reasonable to conclude that no mathematical situation would have occurred within the first five minutes. Because the study aimed at investigating the experience of MA during math classes (and not the prospective trait-MA beforehand), it was important to exclude potential phases not tied to mathematics from the analysis. Furthermore, an initial level of (trait-)MA as assessed during the first minutes of class might have biased the results of the ESM assessments during the lessons. The remaining 40 minutes of the lessons were divided into four 10-minute intervals. Subsequently, one signal was randomly allocated to each interval in advance.

Prior to the ESM survey, the students were provided with a comprehensive briefing on the methodology and the specific instructions for completing the questionnaire. The briefing encompassed fundamental information pertaining to the classification and nomenclature of emotional states, including the differentiation between positive and negative affect. Furthermore, the students were provided with instructions on how to complete the questionnaire during the course of the lessons. In particular, they were instructed to provide a report of their current experiences during the lesson. Following the study, students were provided with a debriefing session, during which they

were thanked for their participation.

The study was settled in Grade 4 for two reasons. First, not entering in the broad discussion on the onset of MA (e.g., Orbach et al., 2019a), empirical studies consistently suggest that state-MA can be assessed validly in this age group. Second, in Germany, there is only one tier for students in Grade 4, whereas from Grade 5 on, students are assigned to different school types based on their academic achievement. Thus, in Grade 4 students from the whole performance range learn together. In order to avoid allocation and peer comparison effects to influence the measurement of MA and securing for a valid assessment of MA, Grade 4 was chosen.

Sample

A total of 72 fourth-grade students participated in the study. The students visited four different classes from four schools in North Rhine-Westphalia (western part of Germany). All students of the classes who obtained informed consent from the parents participated, which equals about 80% of the invited students. Prior to the data analysis, 15 students were excluded from the study, as their response rate to the signals was below 80% (which corresponds with the average compliance rate of ESM paper-pencil studies as reported by Rintala et al., 2019). The remaining 57 students (24 girls, $M_{\text{age}} = 9.82$ years, $SD_{\text{age}} = .047$ years) responded, on average, to 96.13% ($SD = 4.9\%$) of the signals.

Instruments and variables

The data were assessed in two ways: Prior to the ESM study, the students' trait-MA and gender were assessed via conventional questionnaire. Secondly, students' subjective and situational experiences were evaluated in a variety of scenarios during mathematics classes using a concise yet comprehensive ESM questionnaire. The ESM questionnaire encompassed both

state-MA and the perceived relevance and difficulty of the situation. All assessments were conducted in a classroom setting using paper-and-pencil tests.

Trait-mathematics anxiety. The assessment of students' trait-MA was conducted with the corresponding subscale of the Mathematics Attitudes and Anxiety Questionnaire for Grades 4 and 5 (MAQ 4-5; Orbach et al., 2019b). The MAQ 4-5 is based on the Fragebogen zur Rechenangst (FRA; Krinzing et al., 2007) and the Math Attitude and Anxiety Questionnaire (MAQ; Thomas & Dowker, 2000). The MAQ 4-5 has been designed for use in classroom settings and comprises three subscales, each of which measures a different aspect of students' attitudes, self-efficacy, and trait-MA. The subscale for trait-MA comprises 14 items pertaining to concerns about potential failure in mathematics (e.g., "How worried are you if you encounter difficulties with mathematics homework?"). Participants are invited to indicate their level of agreement with each item on a five-point Likert scale, ranging from "very worried" to "very relaxed". Higher total scores on the MAQ 4-5 indicate a greater intensity of trait-MA. Orbach et al. (2019) report an internal consistency (Cronbach's α) of the trait-MA subscale of the MAQ 4-5 of $\alpha = .92$.

State-mathematics anxiety. The state-MA of the students was evaluated using a six-item scale designed to assess situational anxiety. The items of the scale were constructed based on a comprehensive review of existing state-MA scales (see Orbach et al., 2019b). The items addressed the students' anxiety in the immediately preceding mathematical situation (e.g., "At this moment, I was nervous"). Students were required to indicate their level of agreement with the statements presented on a five-point Likert scale. The internal consistency of the items in this study was deemed to be acceptable ($\alpha = .79$).

Relevance of the situation. The relevance of the situation was evaluated using three items that were developed based on the-

oretical considerations (e.g., Prinski et al., 2018). The items assessed the individual's perception of the task's importance in general, the level of effort exerted (addressing motivation to attain good performance in the classroom, e.g., "I gave my best"), and the tolerance for errors (addressing importance for goals or values, e.g., "I must not make any mistakes"). As with state-MA, students indicated their level of agreement with given statements on a five-point Likert scale. The internal consistency of the scale in this study was found to be acceptable ($\alpha = .77$).

Difficulty of current tasks. Students were asked to indicate the difficulty of the tasks they were currently engaged with one item („How difficult was what you have been doing right now?“). The self-developed seven-point Likert scale was labelled at the endpoints and at the midpoint. The endpoints of the scale were labelled „too easy“ and „too hard“, while the midpoint was labelled „just right“. Single items are typically used in ESM studies to keep the questionnaire short and simple (Fritz et al., 2024).

Evidence supports for the use of single-item measures in ESM studies (Song et al., 2023).

Gender. Student's gender was obtained from the lists of students that schools are obliged to keep. When enrolling in school, usually the gender for each student is recorded from their passports.

Data analyses

In ESM studies, it is of crucial importance to take the multilevel data structure into account, as several measurement occasions are nested within different individuals. Therefore, we adopted a multi-level regression modeling approach. More specifically, we employed a *random-intercept-and-slope* model (for an example see Venetz & Zurbruggen, 2015) both for situation relevance as well as task difficulty individually. First, situational relevance and perceived task difficulty were modelled as predictors of state-MA at the signal level (RQ1; see Figure 2). Following the state-trait-model of MA (Lazarus, 2001; Orbach et al., 2019b),

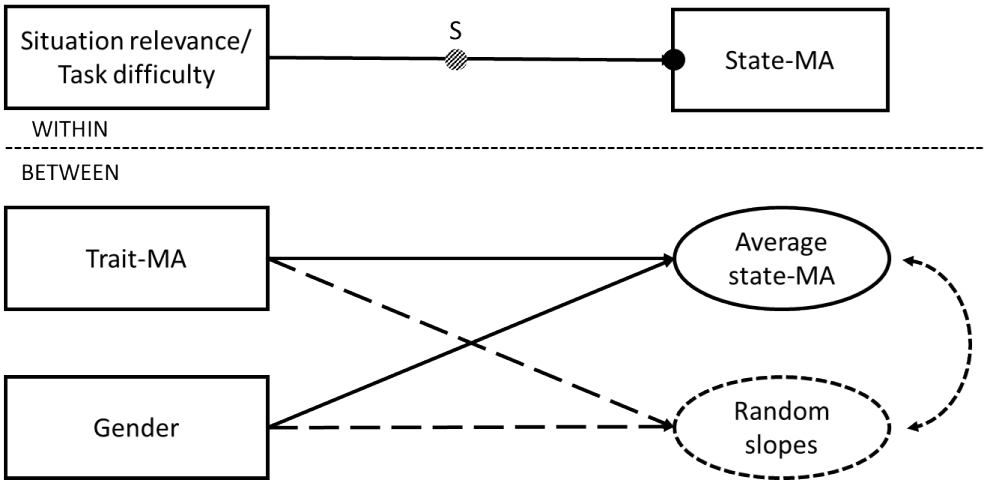


Figure 2
Graphical presentation of the employed intercept-and-slope-as-outcome models

* Note. The points in the within-level represent the random slopes (hatched) and the average state-MA (bold). Solid lines represent model 1, short dashed lines and frames as well as hatched points represent model 2, and long dashed lines represent model 3.

the individual appraisal of the situation relevance and mathematical resources to cope with the situation are distinct processes that happen in a timely manner. Moreover, teachers can address both situational characteristics in different ways, for instance by selecting more or less difficult tasks or embedding them in a more (or less) realistic context. For these reasons, relevance and difficulty were analyzed in two different models to investigate their effect on perceived MA.

At the person level, trait-MA was included as a control variable to assess a general individual tendency towards MA. To investigate for potential gender differences (e.g., higher state-MA in girls; Orbach, 2019a) both in the level of experienced state-MA as well as moderator for the effect of difficulty and relevance, gender was included as a person-level predictor variable for state-MA. In light of the substantial variance assumed for the respective predictors, random intercepts were incorporated in a second step. To address potential differences in the effects of situational relevance and task difficulty on state-MA (RQ2), interaction effects between the predictors and gender, as well as trait-MA, were incorporated in the third and final step.

Results

Descriptive Statistics

Table 1 presents an overview of the descriptive statistics. The results indicated that girls reported a higher degree of state-MA, whereas boys reported slightly higher trait-MA. The female participants rated the relevance and difficulty of the current situation as higher than the male participants. While not all t-tests reached significance at the individual level, the effect sizes were notable for state-MA, situational relevance, and task difficulty. However, the gender differences regarding trait-MA were found to have a very low effect size. At the signal level,

state-MA demonstrated a substantial variance, with a mean state-MA ranging from 0 to 1.99, and a slightly lower range for girls than for boys.

Multi-level regression models

Prior to the regression analyses, an intercept-only model was employed to obtain the intraclass coefficient (ICC). The ICC of .517 indicates that approximately 51.7% of the variance in state-MA can be attributed to inter-individual differences. Consequently, 48.3% of the variance in state-MA can be ascribed to intra-individual differences between the various situations.

Table 2 provides an overview of the multi-level models and the associated coefficients. The relevance of the situation was found to have no significant direct effect on state-MA. However, the difficulty of the current tasks was positively associated with state-MA. This indicates that tasks perceived as more difficult resulted in higher state-MA. Neither situational relevance nor task difficulty demonstrated a significant interaction effect with trait-MA. However, the relevance of the situation demonstrated a positive interaction effect with gender on state-MA. As gender was coded with higher values for girls, the interaction can be interpreted as indicating a greater impact of situational relevance on state-MA in girls than in boys. No interaction effects were identified with regard to task difficulty.

It is notable that the ICCs differ between the models for situational relevance and task difficulty. The substantially higher ICCs in the models regarding the relevance of the situation indicate that the effect of this predictor differed more strongly between participants, as more variance of state-MA was explained by inter-personal variance. In contrast, the ICCs of the models regarding task difficulty were lower and similar to the intercept-only model, which suggests that the effect of task difficulty was more consistent across participants.

Table 1
Descriptive statistics of the measures

	Full sample		Girls		Boys		<i>t</i> (56)	Cohen's <i>d</i>
	<i>M</i> (<i>SD</i>)	Min.- Max.	<i>M</i> (<i>SD</i>)	Min.- Max.	<i>M</i> (<i>SD</i>)	Min.- Max.		
State-MA	.67 (.60)	0- 1.99	.77 (.50)	0- 1.78	.58 (.66)	0- 1.99	1.201	.32
Trait-MA	31.11 (8.92)	6- 48	30.64 (9.86)	6- 48	31.48 (8.07)	12- 45	-.338	-.09
Relevance of situation	2.17 (1.15)	0- 3.86	2.39 (1.01)	0- 3.73	1.99 (1.25)	0- 3.86	1.348	.36
Difficulty of tasks	1.73 (1.09)	0- 4.03	1.97 (1.09)	0- 4.03	1.53 (1.06)	0- 3.11	1.539	.42

**Note.* MA = mathematics anxiety; *M* = mean; *SD* = standard deviation

Table 2
Standardized estimates of the multi-level regression models (random-intercept-and-slope-as-outcome) for the relevance of the situation (top section) and task difficulty (bottom section)

	Model 1		Model 2		Model 3	
	<i>b</i>	<i>SE</i>	<i>b</i>	<i>SE</i>	<i>b</i>	<i>SE</i>
<i>Relevance of situation</i>						
Intercept	.640*	.305	.483	.317	.433	.471
Relevance	-.011	.017	.058	.045	.098	.559
Trait-MA	-.001	.009	.003	.009	.011	.168
Gender	.217	.162	.001	.156	-.439	.256
Relevance x trait-MA					-.004	.005
Relevance x gender					.200*	.089
Variance random slopes			.092***	.036	.086***	.034
Covariance			-.207	.168	-.197	.167
ICC	.517		.621		.617	
<i>Task difficulty</i>						
Intercept	.260	.279	.249	.273	.195	.286
Difficulty	.140***	.010	.128***	.023	.206*	.093
Trait-MA	.003	.008	.005	.008	.008	.008
Gender	.158	.148	.049	.144	-.003	.150
Difficulty x trait-MA					-.004	.003
Difficulty x gender					.071	.045
Variance random slopes			.020	.015	.019	.014
Covariance			-.012	.063	-.011	.064
ICC	.494		.552		.551	

** Note.* MA = mathematics anxiety; *b* = estimates; *SE* = standard error; ICC = intraclass coefficient; * *p*<.05; ** *p*<.01; *** *p*<.001.

In light of the effect of situational relevance on girls was found to be positive and significant ($b = .595$, $SE = .191$, $p = .014$), whereas the same effect on boys was negative but not significant ($b = -.281$, $SE = .241$, $p < .142$). This indicates that girls exhibited heightened state-MA in situations they deemed more relevant, whereas boys demonstrated decreased state-MA in situations deemed more relevant to them. The ICCs and random slope variance exhibited notable discrepancies between the two genders, with lower ICC and higher random slope variance observed in girls (ICC = .448, random slope variance = .110) compared to boys (ICC = .720, random slope variance = .053).

Discussion

This study aimed at investigating the intra- and inter-individual variance of the effect of situational characteristics such as situation relevance and task difficulty on state-MA (RQ1) as well as gender differences in this interplay (RQ2). In order to account for intra-individual variance, ESM were employed. The results of our study provide evidence that ESM can be applied to the assessment of state-MA. In particular, intra- and inter-individual variance in students' experience of state-MA was sufficient for appropriate statistical analyses such as multilevel models.

The role of situational characteristics for state-mathematics anxiety

Our findings offer valuable insights into the effects of perceived situational characteristics on students' experience of MA (RQ1). Overall, the relevance of the situation had no significant effect on students' MA. Given the more consistent, yet counteracting effects of relevance when analyzing only boys or girls, the not existing overall effect might stem from a counterbalance of the gender specific effects. The effects of the

situation relevance on state-MA displayed a substantial intra-individual variability and appeared to be strongly influenced by individual factors as reflected in lower model ICCs. Thus, proximal factors such as mood or social interactions within the classroom might moderate the effects of relevance on MA considerably. Task difficulty, however, was associated with higher levels of MA in all students. Regardless of gender, more difficult tasks were consistently associated with increased MA. This finding suggests that students react to tasks they find difficult in a similar way.

Against the background of the state-trait-model of mathematics anxiety (Orbach et al., 2019b), the impact of the first appraisal (relevance of the situation) on the emergence of MA might be considered smaller and more individual, compared to the second appraisal (difficulty of the task). The fact that students' perception of situational characteristics account for nearly half of the variance in students' emotional responses underscores the importance of contextual factors in shaping experiences of MA (Ramirez et al., 2018). The second appraisal, which relates to task difficulty, seems to have a more predictable and uniform impact compared to the first appraisal. This result highlights the importance of students' individual perception to mathematical situations for their experience of MA. Substantial ICCs indicate that the intra-individual variance and the perception (Ramirez et al., 2018).

Gender differences regarding situational factors

With regard to gender differences in the interplay of situation relevance and task difficulty (RQ2), our results reveal both similarities and differences between girls and boys in how situational characteristics affect their experience of MA. Girls and boys demonstrated similar responses to task difficulty (second appraisal), reinforcing the idea that perceived difficulty triggers comparable

emotional reactions across students. This suggests that perceived task difficulty is appraised in a consistent manner, independent of gender.

However, significant gender differences were found in the way boys and girls appraise the relevance of the situation (first appraisal). Our findings showed more individual variance in random slopes as well as higher ICCs for girls than for boys, suggesting that girls interpret the relevance of tasks more heterogeneously than boys, who displayed a more uniform pattern. For girls, the relevance of the task seemed more closely tied to their concerns about performing well and meeting academic expectations, which is consistent with research suggesting that girls are more likely to be driven by fears of failure in academic contexts (Orbach et al., 2019b). In contrast, boys may place greater emphasis on peer status (e.g., being perceived as competent; Rose & Rudolph, 2006), which can lead to a different appraisal of the task's relevance. This divergence in appraisal may reflect underlying differences in motivation and social dynamics, with girls possibly focusing more on personal success and academic achievement, while boys might prioritize their social standing among peers (e.g., Legewie & DiPrete, 2012).

Interestingly, we found no significant gender differences in trait-MA, which contrasts with existing literature (Hill et al., 2016; Krinzinger et al., 2007; Orbach et al., 2019b). However, girls exhibited higher levels of state-MA in specific situations, suggesting that while their general disposition toward (trait-)MA may be comparable to boys, their emotional reactions to specific (state-)MA are more pronounced (Ashcraft, 2002). This inversion of a frequently reported pattern warrants further investigation and could be indicative for differences in how boys and girls process the immediate stressors of mathematics-related situations, potentially mediated by social stereotypes (Beilock et al., 2012).

Limitations and future directions

At least six limitations of this study must be acknowledged. First, not all students in the sample exhibited high levels of MA, limiting the generalizability of our findings to highly mathematics-anxious populations. Moreover, students with low to no state-MA are a potential source of noise in the data, as their responses might be addressing the mathematical situation to a smaller degree. Future research should focus more explicitly on students with elevated levels of MA to better understand how task difficulty and relevance influence their emotional responses.

Second, we employed a semi-random sampling scheme. In doing so, we did not explicitly assess the specific events that triggered students' emotional responses. An event-contingent sampling scheme, where participants are asked to report when a specific and rather rare event occurs (e.g., a panic attack), could yield clearer insights into when and how students experience particularly anxiety-inducing moments (Dejonckheere & Erbas, 2021). In addition, the tasks students worked on during the assessments were not captured. Thus, it would be worthwhile to assess the tasks worked on in classroom and to evaluate their general difficulty. As an impulse for subsequent studies, the tasks and their difficulty could be systematically manipulated in an experimental approach.

Third, the students' mathematical skills were not included in the study. The second appraisal of the state-trait-model of mathematics anxiety addresses students' coping resources for mathematical situations. The coping resources depend, as operationalized in this study, on difficulty of the tasks, but also on the students' mathematics skills (Orbach et al., 2019b). Integrating both contextual and individual aspects of the second appraisal should provide deeper insights in their interplay. To account for both the actual and the perceived mathematical skills, objective (e.g., standardized tests) and sub-

jective (e.g., math self-concept) measures of mathematics skills are supposed to be added.

Fourth, the classroom setting in which the data were collected may have affected students' appraisals of situation relevance. On the one hand, students might have perceived the classroom situations more relevant in general, because the classroom situations might have an effect on teachers' grading in general. On the other hand, students might have evaluated the situation relevance in relation to other school-based situations, such as exams. In this case, students might have perceived the situations assessed as less relevant. A more targeted assessment of situational relevance (e.g., "Do you think your grade will depend on the current tasks?") could shed more light in this regard.

Fifth, the classroom constitutes a third level in the data structure (i.e., signals are nested in students, and students are nested in classes). For instance, the way teachers give feedback to their students, and in particular how they address mistakes, is supposed to have a substantial influence on both the students' perception of mathematical relevance as well as their mathematics skills (Burns et al., 2021). Moreover, peer status, which might be bound to teacher feedback, as well as social references regarding performance can have substantial effect on state-MA. Thus, future studies might account for class level as it might systematically affect students' second appraisal of MA.

Sixth, this study only covered students in Grade 4 in elementary schools, were students of all performance levels learn together. As for all studies on MA, it is not clear how the results can be generalized on other age groups and school types. Further research on other age ranges or school types with larger sample size on individual and classroom level is necessary.

In conclusion, our study contributes to a deeper understanding of how situational characteristics—particularly task difficulty and relevance—affect students' experienc-

es of MA. Task difficulty appears to have a universal impact, while the relevance of the situation is more subjective and influenced by gender-specific concerns. These findings can inform educational practices, especially in how mathematics tasks are presented to students.

References

- Ashcraft, M. H. (2002). Math anxiety: Personal, educational, and cognitive consequences. *Current Directions in Psychological Science*, 11(5), 181–185. <https://doi.org/10.1111/1467-8721.00196>
- Barroso, C., Ganley, C. M., McGraw, A. L., Geer, E. A., Hart, S. A., & Daucourt, M. C. (2021). A meta-analysis of the relation between math anxiety and math achievement. *Psychological Bulletin*, 147(2), 134–168. <https://doi.org/10.1037/bul0000307>
- Beilock, S. L., Rydell, R. J., & McConnell, A. R. (2007). Stereotype threat and working memory: Mechanisms, alleviation, and spillover. *Journal of Experimental Psychology: General*, 136(2), 256–276. <https://doi.org/10.1037/0096-3445.136.2.256>
- Bolger, N., & Laurenceau, J.-P. (2013). *Intensive longitudinal methods. An introduction to diary and experience sampling research*. Guilford Press.
- Burns, E. C., Martin, A. J., & Evans, P. A. (2021). The role of teacher feedback–feedforward and personal best goal setting in students' mathematics achievement: A goal setting theory perspective. *Educational Psychology*, 41(7), 825–843. <https://doi.org/10.1080/01443410.2019.1662889>
- Carey, E., Devine, A., Hill, F., & Szűcs, D. (2017). Differentiating anxiety forms and their role in academic performance from primary to secondary school. *PLOS ONE*, 12(3), e0174418. <https://doi.org/10.1371/journal.pone.0174418>
- Dejonckheere, E., & Erbas, Y. (2021). Designing an experience sampling study. In I. Myin-Germeyns & P. Kuppens (Eds.), *The open handbook of experience sampling methodology: A step-by-step guide to designing, conducting, and analyzing ESM studies* (pp. 33–70). Center for Research on Experience Sampling and Ambulatory Methods Leuven.

- Devine, A., Fawcett, K., Szűcs, D., & Dowker, A. (2012). Gender differences in mathematics anxiety and the relation to mathematics performance while controlling for test anxiety. *Behavioral and Brain Functions*, 8(1), Art. 33. <https://doi.org/10.1186/1744-9081-8-33>
- Dowker, A., Bennett, K., & Smith, L. (2012). Attitudes to mathematics in primary school children. *Child Development Research*, 2012, 124939. <https://doi.org/10.1155/2012/124939>
- Dowker, A., Sarkar, A., & Looi, C. Y. (2016). Mathematics anxiety: What have we learned in 60 years? *Frontiers in Psychology*, 7, 508. <https://doi.org/10.3389/fpsyg.2016.00508>
- Foley, A. E., Herts, J. B., Borgonovi, F., Guerriero, S., Levine, S. C., & Beilock, S. L. (2017). The math anxiety-performance link: A global phenomenon. *Current Directions in Psychological Science*, 26(1), 52–58. <https://doi.org/10.1177/0963721416672463>
- Fritz, J., Piccirillo, M. L., Cohen, Z. D., Frumkin, M., Kirtley, O., Moeller, J., Neubauer, A. B., Norris, L. A., Schuurman, N. K., Snippe, E., & Bringmann, L. F. (2024). So you want to do ESM? 10 essential topics for implementing the experience-sampling method. *Advances in Methods and Practices in Psychological Science*, 7(3), Article 25152459241267912. <https://doi.org/10.1177/25152459241267912>
- Hektner, J. M., Schmidt, J. A., & Csikszentmihalyi, M. (2007). *Experience sampling method*. SAGE. <https://doi.org/10.4135/9781412984201>
- Hill, F., Mammarella, I. C., Devine, A., Caviola, S., Passolunghi, M. C., & Szűcs, D. (2016). Maths anxiety in primary and secondary school students: Gender differences, developmental changes and anxiety specificity. *Learning and Individual Differences*, 48, 45–53. <https://doi.org/10.1016/j.lindif.2016.02.006>
- Krinzinger, H., Kaufmann, L., Dowker, A., Thomas, G., Graf, M., Nuerk, H. C., & Willmes, K. (2007). Deutschsprachige Version des Fragebogens für Rechenangst (FRA) für 6- bis 9-jährige Kinder [German version of the math anxiety questionnaire (FRA) for 6- to 9-year-old children]. *Zeitschrift für Kinder- und Jugendpsychiatrie und Psychotherapie*, 35(5), 341–351. <https://doi.org/10.1024/1422-4917.35.5.341>
- Lazarus, R. S. (2001). Relational meaning and discrete emotions. In K. R. Scherer, A. Schorr, & T. Johnstone (Eds.), *Appraisal processes in emotion: Theory, methods, research* (pp. 37–67). Oxford University Press.
- Legewie, J., & DiPrete, T. A. (2012). School Context and the Gender Gap in Educational Achievement. *American Sociological Review*, 77(3), 463–485. <https://doi.org/10.1177/0003122412440802>
- Maloney, E. A., Waechter, S., Risko, E. F., & Fugelsang, J. A. (2012). Reducing the sex difference in math anxiety: The role of spatial processing ability. *Learning and Individual Differences*, 22(3), 380–384. <https://doi.org/10.1016/j.lindif.2012.01.001>
- Moran, T. P. (2016). Anxiety and working memory capacity: A meta-analysis and narrative review. *Psychological Bulletin*, 142(8), 831–864. <https://doi.org/10.1037/bul0000051>
- Namkung, J. M., Peng, P., & Lin, X. (2019). The relation between mathematics anxiety and mathematics performance among school-aged students: A meta-analysis. *Review of Educational Research*, 89(3), 459–496. <https://doi.org/10.3102/0034654319843494>
- OECD. (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. OECD Publishing. <https://doi.org/10.1787/9789264190511-en>
- O'Leary, K., Fitzpatrick, C. L., & Hallett, D. (2017). Math anxiety is related to some, but not all, experiences with math. *Frontiers in psychology*, 8, 2067
- Orbach, L., Herzog, M., & Fritz, A. (2019a). Math anxiety during the transition from primary to secondary school. In M. Knigge, D. Kollasche, O. Skovsmose, R. M. J. de Souza, & M. G. Penteado (Eds.), *Inclusive mathematics education: Research results from Brazil and Germany* (pp. 419–448). Springer.
- Orbach, L., Herzog, M., & Fritz, A. (2019b). Relation of state- and trait-math anxiety to intelligence, math achievement, and learning motivation. *Journal of Numerical Cognition*, 5(3), 371–399. <https://doi.org/10.5964/jnc.v5i3.252>
- Orbach, L., Herzog, M., & Fritz, A. (2020). State- and trait-math anxiety and their relation to math performance in children: The role of core executive functions. *Cognition*, 200, 104271. <https://doi.org/10.1016/j.cognition.2020.104271>

- Orbach, L., Schmitz, F., & Fritz, A. (2025). Gender Differences and the Latent Structure of Questionnaires in Math Anxiety in Children. *Lernen und Lernstörungen* (online first). <https://doi.org/10.1024/2235-0977/a000485>
- Richardson, F. C., & Suinn, R. M. (1972). The mathematics anxiety rating scale: psychometric data. *Journal of counseling Psychology, 19*(6), 551–554.
- Rintala, A., Wampers, M., Myin-Germeys, I., & Viechtbauer, W. (2019). Response compliance and predictors thereof in studies using the experience sampling method. *Psychological Assessment, 31*(2), 226–235.
- Rose, Amanda J.; Rudolph, Karen D. . (2006). A Review of Sex Differences in Peer Relationship Processes: Potential Trade-offs for the Emotional and Behavioral Development of Girls and Boys. *Psychological Bulletin, 132*(1), 98–131. <https://doi.org/10.1037/0033-2909.132.1.98>
- Schmidt, P., Jendryczko, D., Zurbriggen, C. L. A., & Nussbeck, F. W. (2023). Recall bias of students' affective experiences in adolescence: The role of personality and internalizing behavior. *Journal of Adolescence, 95*, 893–906. <https://doi.org/10.1002/jad.12162>
- Song J., Howe E., Oltmanns J. R., Fisher A. J. (2023). Examining the concurrent and predictive validity of single items in ecological momentary assessments. *Assessment, 30*(5), 1662–1671. <https://doi.org/10.1177/10731911221113563>
- Sorvo, R., Koponen, T., Viholainen, H., Aro, T., Raikkonen, E., Peura, P., & Aro, M. (2017). Math anxiety and its relationship with basic arithmetic skills among primary school children. *British Journal of Educational Psychology, 87*(3), 309–327. <https://doi.org/10.1111/bjep.12151>
- Spielberger, C. D. (1972). *Anxiety. Current trends in theory and research*. Academic Press.
- Thomas, G., & Dowker, A. (2000). Mathematics anxiety and related factors in young children. Paper presented at the British Psychological Society Developmental Section Conference, Bristol, UK.
- Venetz, M. & Zurbriggen, C. (2015). Intensive longitudinal methods – ihre Eignung für die sonderpädagogische Forschung und exemplarische Anwendungsmöglichkeiten. *Empirische Sonderpädagogik, 7*(3), 194–205.
- Wood, G., Pinheiro-Chagas, P., Julio-Costa, A., Micheli, L., Krinzinger, H., Kaufmann, L., & Haase, V. G. (2012). Math anxiety questionnaire: Similar latent structure in Brazilian and German school children. *Child Development Research, 2012*, 10. <https://doi.org/10.1155/2012/610849>
- Wu, S. S., Willcutt, E. G., Escovar, E., & Menon, V. (2014). Mathematics achievement and anxiety and their relation to internalizing and externalizing behaviors. *Journal of Learning Disabilities, 47*(6), 503–514. <https://doi.org/10.1177/0022219413478850>
- Zhang, J., Zhao, N., & Kong, Q. P. (2019). The relationship between math anxiety and math performance: A meta-analytic investigation. *Frontiers in Psychology, 10*, 1613. <https://doi.org/10.3389/fpsyg.2019.01613>
- Zurbriggen, C. L. A., Jendryczko, D., & Nussbeck, F. W. (2021). Rosy or blue? Change in recall bias of students' affective experiences during early adolescence. *Emotion, 21*(8), 1637–1649. <https://doi.org/10.1037/emo0001031>

Author Note

Corresponding author

 Moritz Herzog
<https://orcid.org/0000-0002-6706-3351>

 Miriam Balt
<https://orcid.org/0000-0001-7842-109X>

 Carmen Zurbriggen
<https://orcid.org/0000-0002-0906-2000>

Moritz Herzog
University of Cologne
Institute for Mathematics Education
Gronewaldstrasse 2
D-50931 Cologne
moritz.herzog@uni-koeln.de

*No conflict of interest is tied to this research.
The study was conducted without any additional funding.*

Erstmals eingereicht: 04.10.2024
Überarbeitung eingereicht: 21.04.2025
Angenommen: 26.06.2025

Offene Daten	Daten können unter moritz.herzog@uni-koeln.de angefragt werden.
Offener Code	Der zur Analyse der Daten verwendete R code kann unter moritz.herzog@uni-koeln.de angefragt werden.
Offene Materialien	Die verwendeten Materialien können unter moritz.herzog@uni-koeln.de angefragt werden.
Präregistrierung	Nein
Votum Ethikkommission	Die Ethikkommission der Bergischen Universität Wuppertal hat am 23.01.2023 das Design der vorliegenden Studie geprüft.
Finanzielle und weitere sachliche Unterstützung	Die Studie wurde ohne finanzielle oder sachliche Unterstützung Dritter durchgeführt.
Autorenschaft	Moritz Herzog: Conceptualization, Methodology, Investigation, Data Curation, Formal Analysis, Writing – Original Draft Preparation, Writing - Review & Editing; Miriam Balt: Conceptualization, Methodology, Investigation, Data Curation, Writing – Original Draft Preparation, Writing - Review & Editing; Carmen Zurbriggen: Conceptualization, Methodology, Formal Analysis, Writing – Original Draft Preparation, Writing - Review & Editing
KI und generative Modelle	Zur Prüfung auf Rechtschreib- und Grammatikfehler wurde DeepL Version 25.1 genutzt. Die Vorschläge wurden von den Autor:innen einzeln gegengeprüft.

Welche Rolle spielt das Fördersetting? Längsschnittstudie zur Wirkung der Förderung in der Grafomotorik

*Eckhart Michael¹, Schwery Klingele Joséphine¹,
Sägesser Wyss Judith¹ & Wüthrich Sergej²*

¹ Institut für Heilpädagogik, Pädagogische Hochschule Bern

² Institut für Forschung, Pädagogische Hochschule Bern

Zusammenfassung

Der vorliegende Beitrag nimmt die laufende Inklusionsdiskussion auf und konzentriert sich auf die Wirkung verschiedener grafomotorischer Fördersettings. Im Fokus der Analysen stehen separative, integrative und inklusive Fördersettings. In einem quasi-experimentellen Prä-Post-Follow-up-Design wurden 1337 SchülerInnen der 1. Klassenstufe aus 87 Schulklassen über einen Zeitraum von 15 Monaten begleitet. Die Schulklassen wurden drei verschiedenen Fördersettings zugeteilt: separativ (n = 24), integrativ (n = 21) und inklusiv (n = 42). Die grafomotorische Leistung wurde mittels des GRAFOS-2 Screenings zu drei Messzeitpunkten erfasst. Alle Fördersettings implementierten eine 16-wöchige Intervention (2 Lektionen/Woche). Die Datenanalyse erfolgte mit hierarchischen gemischten Modellen und Propensity Score Matching. Die Ergebnisse zeigen, dass SchülerInnen in allen Fördersettings vergleichbare Fortschritte in ihren grafomotorischen Leistungen erzielten. Bei Betrachtung der SchülerInnen mit ausgewiesenem grafomotorischem Förderbedarf im separativen Setting sprechen die längerfristigen Effekte jedoch eher für integrative und inklusive Förderangebote.

Schlagwörter: Grafomotorik, Fördersettings, Separation, Integration, Inklusion

The Role of the Intervention Setting: A Longitudinal Study on the Effects of Support in Graphomotor Skills

Summary

This article takes up the ongoing discussion on inclusion and focuses on the effect of different graphomotor support settings. The analyses focus on separative, integrative and inclusive graphomotor support settings. In a quasi-experimental pre-post follow-up design, 1,337 first graders from 87 school classes were followed over a period of 15 months. The classes were assigned to three different support settings: separate (n = 24), integrated (n = 21) and inclusive (n = 42). Graphomotor performance was assessed using the GRAFOS-2 screening at three time points. All settings implemented a 16-week intervention (2 lessons/week). Data were analysed using hierarchical mixed models and propensity score matching. The results show that students in all intervention settings made comparable progress in

their graphomotor skills. However, when considering students with identified graphomotor needs in the separate setting, the longer term effects tended to favour integrative and inclusive support programmes.

Keywords: graphomotor skills, intervention setting, segregation, integration, inclusion

In den letzten Jahrzehnten hat die Forderung nach schulischer Inklusion an Bedeutung gewonnen. Inzwischen haben über 180 Länder die UN-Behindertenrechtskonvention unterzeichnet (UNO, 2006), die ein inklusives Bildungssystem auf allen Ebenen proklamiert (vgl. Art.24). Auch pädagogisch-therapeutische Angebote, insbesondere die Psychomotorik und Logopädie, sind gefordert, ihre Unterstützungsmassnahmen im Sinne inklusiver Bildung weiterzuentwickeln. Im Bereich der Psychomotorik gehört das therapeutische Einzel- und Kleingruppensetting zur klassischen Unterstützungsform (Vetter et al., 2009). In den vergangenen Jahren wurden jedoch wenige Förderkonzepte und -materialien entwickelt, die darauf abzielen, die bislang separativ organisierte Psychomotoriktherapie in ein inklusives Schulsetting zu integrieren (Sägesser Wyss et al., 2021a; Vetter et al., 2021). Diese Entwicklung bedingt eine Umgestaltung des traditionellen Handschriftunterrichts, sodass der Unterricht auf die heterogenen Bedürfnisse aller SchülerInnen abgestimmt werden kann (Sägesser Wyss et al., 2021b). Der folgende Beitrag nimmt die skizzierten Entwicklungen auf und bearbeitet die Fragestellung nach den Wirkungen separativer, integrativer und inklusiver Fördersettings auf grafomotorische Kompetenzen.

Grafomotorische Kompetenzen

Grafomotorische Fertigkeiten umfassen die Vorläuferfertigkeiten des Schreibens von Hand sowie den Transfer dieser Vorläuferfertigkeiten in das Schreiben von Buchstabenformen (Sägesser Wyss et al., 2021a). Als zentrale Prädiktoren der Grafomotorik gelten die visuomotorische Integration und

die Feinmotorik (Achymy et al., 2022). Die visuomotorische Integration beschreibt die Fähigkeit zur Reproduktion von Formen, welche auf dem Zusammenspiel der visuellen Wahrnehmung, der mentalen Repräsentation einer Form und der Motorik des Arms basiert (Beery et al., 2010). Die Feinmotorik hingegen bezieht sich auf die präzise Koordination und Kontrolle der Auge-, Hand- oder Fingermuskulatur (Chandler et al., 2021). Die Grafomotorik ist ein Teilbereich der Psychomotorik und steht in enger Verbindung mit der sensorischen, motorischen, sozio-emotionalen und kognitiven Entwicklung (Sägesser Wyss et al., 2024). Grafomotorische Kompetenzen bilden eine zentrale Grundlage für den Schriftspracherwerb (Kranz et al., 2011).

Einordnung der Grafomotorik in den Schriftspracherwerb

Um die Einbettung der Grafomotorik in den Schriftspracherwerb zu erklären, wird das Schreiben als ein komplexer Prozess verstanden, der im Modell von Van Galen (1991) als hierarchisch organisierte Ebenen dargestellt wird. In diesem Modell werden grafomotorische Fertigkeiten auf der motorischen Ebene verortet und umfassen das Schreiben von Schriftsymbolen (Formen und Buchstaben). Darauf aufbauend folgt die handschriftliche Ebene, die das Schreiben von Wörtern und die Auswahl der Wörter in der richtigen Reihenfolge einschliesst, um grammatikalisch korrekte und sinnvolle Sätze zu bilden. Abschliessend folgt die intentionale Ebene, in der eine Textidee entwickelt wird (Van Galen, 1991). Neben der hierarchischen Darstellung unterschiedlicher und aufeinander aufbauender Ebenen, die im Schreibprozess involviert sind, zeigt

van Galen (1991) die gleichzeitige Beteiligung kognitiver Aspekte auf allen Ebenen.

Relevanz der Grafomotorik

Die hierarchische Struktur des Schreibprozesses wurde in zahlreichen Untersuchungen bestätigt (z. B. Salas & Silvente, 2020). So konnte gezeigt werden, dass grafomotorische Fertigkeiten die Handschriftentwicklung voraussagen (z. B. Duiser et al., 2014; Maurer et al., 2023) und die Textproduktion beeinflussen (z. B. Skar et al., 2021). In diesem Zusammenhang ist jedoch zu beachten, dass die verschiedenen Ebenen nicht klar voneinander zu trennen sind, sondern dass es vielmehr zu Überschneidungen bzw. Überlappungen zwischen den verschiedenen Teilprozessen des Schreibens kommt (Behrens, 2017). Berninger und Winn (2006) stellen ein Schreibmodell vor, das sich in Abgrenzung zum Modell von Van Galen (1991) weniger auf die hierarchische Anordnung von Prozessen konzentriert, sondern vielmehr die Komplexität der beim Schreiben gleichzeitig ablaufenden Teilprozesse verdeutlicht (Berninger et al., 2015). In diesem Kontext spielt das Arbeitsgedächtnis eine zentrale Rolle (Berninger, 2006). In den frühen Stadien des Schreibens ist das visuell-räumliche Arbeitsgedächtnis besonders wichtig, um sich die auf der Vorlage abgebildete geometrische Form oder Buchstabenform zu merken und zu behalten und die Ausführung dieser Form zu planen (Maurer et al., 2023). Kinder, die das Schreiben von Buchstaben erlernen, müssen die Bewegungsabläufe noch bewusst steuern, was einen beträchtlichen Teil der begrenzten Kapazität des Arbeitsgedächtnisses beansprucht (McCutchen, 2011). Erst wenn die grafomotorischen Prozesse automatisiert sind, werden Kapazitäten im Arbeitsgedächtnis frei (Becker et al., 2014), was wiederum zu einer besseren Schreibleistung führen sollte (Sturm et al., 2017).

Vor diesem Hintergrund ist es von Bedeutung, grafomotorische Fertigkeiten bereits im Vorschulalter auf kindgerechte Weise zu fördern (Vetter et al., 2021) und diese Förderung im frühen Schulalter gezielt fortzusetzen, um die Bewegungsabläufe weitgehend zu automatisieren und optimale Bedingungen für den weiteren schulischen Lernprozess zu schaffen (Kushki et al. 2011; Maurer et al., 2023). Die Folgen unerkannter grafomotorischer Schwierigkeiten können für die betroffenen Kinder nachhaltig negativ sein. Es wird von negativen Auswirkungen auf die Motivation, das Selbstvertrauen und der Verlauf der Schullaufbahn berichtet (Dinehart, 2014; Eckhart & Sägeser Wyss, 2016).

Von der Separation zur Inklusion

Die traditionelle Sonderpädagogik ist durch die Separation von *SchülerInnen mit sonderpädagogischem Förderbedarf*¹ von ihren MitschülerInnen ohne Förderbedarf gekennzeichnet. Diese Trennung erfolgt in der Regel in separaten, häufig kleineren Klassen oder in speziellen Bildungseinrichtungen. In den vergangenen Jahrzehnten ist diese separative Förderung zunehmend in Kritik geraten. Die in diesem Zusammenhang entstandene Integrationsbewegung zielte darauf ab, die schulische Separation schrittweise aufzulösen und SchülerInnen mit sonderpädagogischem Förderbedarf in das allgemeine Bildungssystem einzubinden (Haeberlin, 2005). Gegenwärtig werden unter dem Begriff der schulischen Inklusion diese Entwicklungen fortgeführt. In der wissenschaftlichen Debatte werden die Begriffe Inklusion und Integration häufig synonym verwendet und basieren auf der Grundannahme, dass die Vielfalt der Lernenden anerkannt und Ausgrenzung sowie Stigmatisierung vermieden werden sollen (Sahli Lozano et al., 2017). Dies impliziert eine veränderte Perspektive auf die Heterogenität der SchülerInnen hin-

¹ Der Begriff *SchülerInnen mit sonderpädagogischem Förderbedarf* umfasst eine äußerst vielfältige Gruppe von Kindern mit einer breiten Palette physischer, kognitiver und sozioemotionaler Schwierigkeiten, die dazu führen, dass sie unterschiedliche Grade an sonderpädagogischem Förderbedarf benötigen (Dalgaard et al., 2022).

sichtlich unterschiedlicher Dimensionen (z.B. Herkunft, Geschlecht, Begabungen, Lernvoraussetzungen etc.) (Prengel, 2019) und erfordert eine Unterrichtsgestaltung, die den individuellen Bedürfnissen aller SchülerInnen gerecht wird (Bless, 2018; Krämer et al., 2021). Die inklusive Schule beschreibt eine Vision, deren Realisierung schrittweise angegangen werden muss (Sahli Lozano et al., 2017). Einen Orientierungsrahmen für die Umsetzung von Inklusion bietet der *Index für Inklusion* (Booth und Ainscow, 2019). Dieser beschreibt drei Dimensionen, die im Entwicklungsprozess berücksichtigt werden müssen: a) inklusive Kulturen (z. B. Haltungen), b) inklusive Strukturen (z. B. Rahmenbedingungen, Konzepte) und c) inklusive Praktiken (z. B. Unterricht, (multi-) professionelle Zusammenarbeit). Die vorliegende Studie konzentriert sich auf die Dimension der inklusiven Praktiken im Kontext des Handschriftunterrichts.

Umsetzung inklusiver Förderansätze im Handschriftunterricht

In der Schweiz erfolgt die Diagnostik und Förderung von SchülerInnen mit Schwierigkeiten in der Handschrift, bzw. in der Grafomotorik in erster Linie durch PsychomotoriktherapeutInnen. Die Psychomotorik ist als fester Bestandteil des sonderpädagogischen Grundangebotes an der Schweizer Volksschule verankert. Obwohl eine Weiterentwicklung psychomotorischer Förderangebote im Sinne einer inklusiven Bildung gefordert wird (vgl. Sahli Lozano et al., 2017), dominieren weiterhin traditionelle Unterstützungsformen in separativen Einzel- und Kleingruppensettings (Vetter et al., 2009). Die Umsetzung inklusiver Förderansätze erfolgte bislang nur selten, sodass sich empirische Befunde zu inklusiven Fördersettings in der Psychomotorik lediglich in Einzelfällen dokumentieren lassen. Ein Beispiel ist die Untersuchung von Kranz et al. (2011), die in der Schweiz durchgeführt wurde. In dieser Studie wurden die Effekte einer psychomotorischen Interven-

tion, *G-Fipps* (Grafomotorische Förderung in integrativ und präventiv ausgerichteter Psychomotorik) (Vetter et al., 2021), auf die grafomotorische Leistungsentwicklung von Kindergartenkindern analysiert. Die Durchführung der Förderstunden erfolgte sowohl durch die Lehrkraft als auch durch die PsychomotoriktherapeutIn. Die Resultate zeigen, dass Kinder der Interventionsgruppe signifikant grössere Fortschritte in Bezug auf grafomotorische Vorläuferfertigkeiten und körperkoordinative Fertigkeiten erzielten als die Kinder in den Kontrollgruppen. Internationale vergleichende Studien zur Wirksamkeit inklusiver grafomotorischer Förderansätze sind ebenfalls selten. Als bedeutend sind in diesem Zusammenhang die Arbeiten von Case-Smith et al. (2012; 2014) zu nennen. In diesen Studien wurde die Förderintervention *Write-Start-Programm* zur Förderung der Handschrift bei ErstklässlerInnen mit und ohne grafomotorischem Förderbedarf eingesetzt. Die Umsetzung der Intervention erfolgte im Rahmen des Co-Teachings durch Lehrkräfte und ErgotherapeutInnen. Die Ergebnisse zeigen signifikante Verbesserungen in den Bereichen Handschrift-Leserlichkeit und Handschrift-Flüssigkeit bei den SchülerInnen der Interventionsgruppen im Vergleich zu den Kontrollgruppen, die regulären Handschriftunterricht erhielten. Eine weitere internationale Studie, die eine grafomotorische Förderkonzeption im Rahmen des Co-Teachings bei Kindergarten- und Erstklasskindern untersuchte, stammt von Taverna et al. (2020). Die Ergebnisse zeigen eine signifikante Verbesserung der visuomotorischen Integration, jedoch nicht der Feinmotorik. Allerdings wurde in dieser Studie keine Kontrollgruppe berücksichtigt, sodass die Aussagekraft hinsichtlich der Wirksamkeit der Intervention limitiert bleibt. Vor dem skizzierten Hintergrund erweist sich die aktuelle Forschungslage zur grafomotorischen Leistungsentwicklung in inklusiven Fördersettings als lückenhaft. Es fehlt insbesondere an systematischen Analysen, die den Einfluss unterschiedlicher grafomotori-

scher Fördersettings auf die grafomotorische Leistungsentwicklung untersuchen. Die vorliegende Studie greift diese Forschungslücke auf, indem sie drei verschiedene Fördersettings untersucht: das separative, das integrative und das inklusive Setting (vgl. Abbildung 1). Das *separative Setting* ist definiert als grafomotorische Förderung, die ausserhalb des regulären Unterrichts durch eine PsychomotoriktherapeutIn erfolgt. Das *integrative Setting* ist dadurch gekennzeichnet, dass die Förderung innerhalb des Klassenzimmers stattfindet, bei der die Psychomotoriktherapeutin gezielt mit den Kindern mit Förderbedarf arbeitet, während die Lehrkraft parallel dazu den Handschriftunterricht für die übrigen SchülerInnen leitet. Das *inklusive Setting* zeichnet sich dadurch aus, dass der Handschriftunterricht von der Lehrkraft und der Psychomotoriktherapeutin gemeinsam nach den Prinzipien einer inklusiven Didaktik (Sägesser Wyss et al., 2021a) für alle Kinder gestaltet wird, unabhängig davon, ob bei einem Kind ein grafomotorischer Förderbedarf vorliegt.

Fragestellungen

Die vorliegende Studie trägt durch die folgenden Forschungsfragen zur evidenzbasierten Analyse der Wirksamkeit unterschiedlicher Fördersettings von SchülerInnen mit und ohne grafomotorischen Förderbedarf bei und liefert Erkenntnisse für die Weiterentwicklung grafomotorischer Fördersettings:

Fragestellung 1: Machen SchülerInnen in integrativen oder inklusiven Settings insgesamt mehr Leistungsfortschritte in der Grafomotorik als die SchülerInnen in separativen Fördersettings?

Fragestellung 2: Machen insbesondere SchülerInnen mit schwacher Grafomotorikleistung², aber ohne ausgewiesenen grafomotorischen Förderbedarf, mehr Leistungsfortschritte in integrativen oder inklusiven Settings im Vergleich zu separativen Fördersettings?

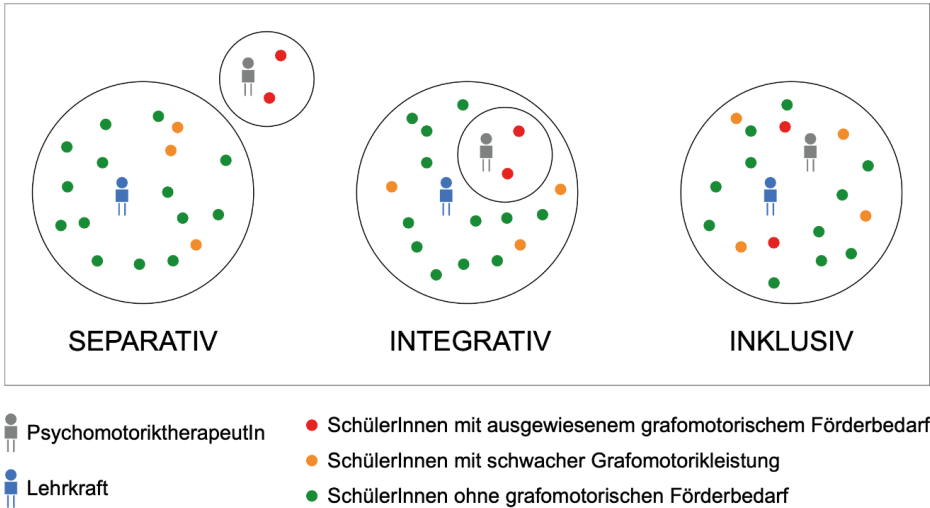


Abbildung 1
Überblick über die drei grafomotorischen Fördersettings

2 SchülerInnen, deren grafomotorische Leistung zum ersten Messzeitpunkt mindestens eine Standardabweichung unter dem Gesamtmittelwert liegen.

Fragestellung 3: Machen SchülerInnen mit ausgewiesenem grafomotorischem Förderbedarf³ in klassenexternen, separativen Therapiesettings mehr Fortschritte in der Grafomotorik als vergleichbare SchülerInnen in integrativen oder inklusiven Fördersettings?

Methoden

Stichprobe

Die Daten entstammen dem nationalen Forschungsprojekt „grafset“ (März 2021–Februar 2024), das von der Ethikkommission der Humanwissenschaftlichen Fakultät der Pädagogischen Hochschule Bern, Schweiz, genehmigt wurde (Genehmigungsnummer: 2020-10-00005). Die Stichprobe umfasste 1337 SchülerInnen (51% Knaben) aus 87 Schulklassen, deren Lehrkräfte ($N = 87$) und PsychomotoriktherapeutInnen ($N = 67$). Die Längsschnittstudie beinhaltete drei Messzeitpunkte: Prä-Test zu Beginn der ersten Klasse (Oktober – Dezember 2021, Durchschnittsalter 7 Jahre, $SD = 0.39$), Post-Test sechs Monate später (April–Juni 2021) und Follow-up zu Beginn der zweiten Klasse (Oktober–Dezember 2022). Die Erziehungsberechtigten wurden im Vorfeld über die Studie informiert und es wurde die Erlaubnis zur Teilnahme der SchülerInnen eingeholt. Die quasi-experimentelle Prä-Post-Follow-up-Studie erfolgte – aufgrund unterschiedlicher Gegebenheiten an den Schulstandorten mit lediglich teilrandomisierter Zuordnung der Schulklassen – in drei Fördersettings: separativ ($n = 24$ Schulklassen), integrativ ($n = 21$) und inklusiv ($n = 42$).-

Intervention

Die Intervention gliederte sich in drei Phasen: *Erstens* erhielten alle Fachpersonen eine zweistündige Weiterbildung zu den

fünf Schwerpunkten der Grafomotorikförderung (visuelle Wahrnehmung, visuomotorische Integration, taktil-kinästhetische Wahrnehmung, Feinmotorik, Körperwahrnehmung/Bewegungsplanung) sowie zu den Charakteristika der drei Fördersettings. Fachpersonen des inklusiven Settings erhielten zusätzlich Materialien zu inklusiver Didaktik und Zusammenarbeit (basierend auf Sägesser Wyss et al., 2021a). *Zweitens* erfolgte die Umsetzung der Förderintervention während 16 Wochen (2 Lektionen/Woche) mit Lektionenprotokollen, begleitet durch zwei Online-Austausche und eine „Hotline“ zum Projektteam zur Sicherstellung konsistenter Umsetzung. In allen Settings wurde an gleichen Förderschwerpunkten gearbeitet, wobei allen Lehrkräften und TherapeutInnen dieselben Fördermaterialien (GRAFINK, Sägesser Wyss et al., 2021a) zur Verfügung standen. *Drittens* wurden nach Projektabschluss alle Beteiligten über zentrale Erkenntnisse informiert, mit optionaler Einführung in inklusive Didaktik und Zusammenarbeit.

Messinstrumente

Grafomotorische Leistung

Zur Erfassung der grafomotorischen Leistung wurde das GRAFOS-2 Screening eingesetzt (Sägesser Wyss et al., 2024). Im Rahmen des GRAFOS-2 Screenings reproduzieren die Kinder acht Grundformen der Schrift (senkrechte Linie, waagerechte Linie, Kreis, Bogen nach unten, Bogen nach oben, Kreuz, Quadrat, Dreieck) sowie vier komplexere Formen (Raute, Doppelschleife, waagerechte Acht, Tropfen). Die präzise Wiedergabe der Formen erfordert visuomotorische Integration und feinmotorische Fertigkeiten. Insgesamt müssen die Kinder jede der 12 Formen jeweils sechs Mal so exakt wie möglich in vordefinierte Felder (1 cm²) reproduzieren. Die Beurteilung erfolgte anhand klar definierter Kriterien. Hierbei

³ SchülerInnen mit ausgewiesenem grafomotorischem Förderbedarf besuchen klassenextern die Psychomotoriktherapie und wurden aufgrund einer Abklärung durch die PsychomotoriktherapeutIn diesem Förderangebot zugewiesen.

wurde jede Form entweder mit 1 (für eine korrekte Wiedergabe) oder mit 0 (für eine ungenaue Wiedergabe) bewertet. Die Interrater-Reliabilität zwischen den verschiedenen Ratern erreichte einen Wert von Fleiss' $\kappa = .65$, was eine gute Übereinstimmung darstellt (Landis & Koch, 1977). Die interne Konsistenz des Screenings, gemessen mit Cronbachs Alpha (α), war ebenfalls hoch ($\alpha = .80$ für die acht Grundformen; $\alpha = .78$ für die vier komplexen Formen).

Demographische Angaben

Die Lehrkräfte beantworteten anhand eines Kurzfragebogens Fragen zu eigenen demographischen Angaben (Alter, Geschlecht) sowie zu den SchülerInnen (Entwicklung Grafomotorik/Motorik, besondere Förderung u.a.m.).

Statistische Analyse

Die Analysen wurden in R (version 4.3.3; R Core Team, 2019) durchgeführt. Zur Bearbeitung der drei Fragestellungen wurden die folgenden Analysen vorgenommen:

1) *Grafomotorische Entwicklung aller SchülerInnen*: Die grafomotorische Leistungsentwicklung wurde mittels hierarchischen gemischten Modellen analysiert, beginnend mit einem Random-Intercept Nullmodell zur Bestimmung des Varianzanteils der Leistung (L1) innerhalb der SchülerInnen (L2) und Schulklassen (L3). Nach Integration der Kontrollvariablen (Alter, Geschlecht) sowie Setting und Zeit als fixe Effekte erfolgte der Vergleich mit einem Random Intercept/Random Slope Modell (Entwicklung über Klassen hinweg zufällig variierend). Im Falle signifikanter Varianz wurde abschliessend die Cross-Level-Interaktion von Setting und Zeit hinzugefügt um zu prüfen, ob sich Entwicklungsunterschiede durch das Setting erklären lassen (Aguinis et al., 2013).

2) *Grafomotorische Entwicklung von SchülerInnen mit schwacher Grafomotorikleistung*: Bei Fragestellung 2 interessierten

Kinder mit schwacher Grafomotorikleistung ohne ausgewiesenen grafomotorischen Förderbedarf, daher wurden SchülerInnen in Therapie ausgeschlossen und jene SchülerInnen selektiert, deren Leistung mindestens eine Standardabweichung unter dem Mittelwert des Restsamples lag ($N = 197$ bzw. 14.7% des Gesamtsamples, vgl. Tabelle 1). Die Analyse erfolgte analog zum Gesamtsample.

3) *Grafomotorische Entwicklung der SchülerInnen mit ausgewiesenem grafomotorischem Förderbedarf*: Der Therapieeffekt im separativen Setting wurde mittels Propensity Score Matching (MatchIt; Ho et al., 2011) untersucht, wobei vergleichbare Gruppen bezüglich Alter, Geschlecht und Baseline-Grafomotorikleistung aus den beiden anderen Settings gebildet wurden (genetisches Matching, $n = 1000$). Die Leistungsentwicklung wurde durch hierarchische gemischte Modelle analysiert, mit reduzierter Komplexität aufgrund geringer Fallzahlen (L1: Leistung zu drei Messzeitpunkten, L2: SchülerInnen; Random Intercept/Fixed Slope Modell).

Zentrierung der Variablen

Die Kontrollvariablen Alter und Geschlecht (Dummy-Variable) wurden am Grand-Mean zentriert, Prä-Test und separatives Setting dienten als Referenzlevels für die Dummy-kodierten Variablen Zeit und Setting.

Umgang mit fehlenden Werten

Während bei Alter und Geschlecht keine fehlenden Werte vorlagen, fehlten vereinzelt Leistungsdaten (Prä-Test: 6.1%, Post-Test: 4.2%, Follow-up: 6.7%), vorwiegend durch krankheitsbedingte Abwesenheiten und unsystematisch (Little's MCAR-Test: $\chi^2 = 10.8$, $df = 9$, $p = .29$). Bei Fragestellung 1 erfolgte deshalb ein einfacher listenweiser Fallausschluss, bei Fragestellung 3 wurde hingegen aufgrund der niedrigen Fallzahlen eine Multiple Imputation durchgeführt (mittels des Pakets MICE; Buuren und Groothuis-

is-Oudshoorn, 2011; mit 100 Imputationen, anschließendem Propensity Score Matching und gepoolten Analyseergebnissen).

Resultate

Deskriptive Resultate

Deskriptive Ergebnisse sind in Tabelle 1 dargestellt. Die Gruppen in den drei Settings unterschieden sich zwar in Bezug auf das Alter, $F(2, 1334) = 9.715, p < .001$, aber nicht in Bezug auf das Geschlecht, $\chi^2(2) = 0.19, p = .91$, oder in Bezug auf die Grafo-motorikleistung zur Prä-Test Messung, $F(2, 81.3) = 0.30, p = .74$. Die unterschiedlichen Gruppengrößen ergeben sich durch die weiteren Fragestellungen, welche im vorliegenden Nationalfondsprojekt bearbeitet werden.

Settingeffekte auf die Entwicklung aller SchülerInnen

Die durchschnittliche Graфомotorikleistung über alle Messzeitpunkte betrug 51.3 Punkte, wobei 28% der Leistungsvarianz durch zufällige Effekte erklärt wurden (20.9% SchülerInnen-L2, 6.7% Klassen-L3). Nach Integration der Kontrollvariablen und Prä-diktoren stieg die erklärte Varianz auf 48.6% (20.1% fixe Effekte), mit Leistungssteigerungen von 9.4 (Post-Test) und 12.9 Punkten (Follow-up). Während Knaben schlechtere (-6.93 Punkte) und ältere SchülerInnen bessere Leistungen (0.14 Punkte/Monat) erzielten, zeigten sich keine signifikanten Unterschiede zwischen den Settings.

Die Hinzunahme von Random Slopes für variierende Leistungsentwicklungen zwischen den Klassen verbesserte den Modellfit signifikant ($\chi^2 = 888.3, df = 5, p < .001$; Post-Test $SD = 9.74$, Follow-up $SD = 11.07$), mit 69.9% erklärter Varianz des Modells. Die Hinzunahme von Cross-Level-Interaktionen des Settings mit der Leistungsentwicklung führte aber zu keiner si-

Tabelle 1
Deskriptive Ergebnisse Stichprobe

Setting	Anzahl N (%)	Alter M (SD)	Geschlecht Junge (%)	Prä-Test M (SD)	Post-Test M (SD)	Follow-up M (SD)
Gesamtstichprobe						
Separativ	369 (27.6)	83.8 (4.74)	188 (50.9)	42.4 (14.4)	54.4 (13.2)	56.1 (14.2)
Integrativ	284 (21.2)	84.3 (4.76)	148 (52.1)	45.0 (13.4)	53.2 (12.8)	56.4 (12.2)
Inklusiv	684 (51.2)	82.9 (4.58)	346 (50.6)	44.8 (13.2)	53.6 (13.5)	57.9 (12.6)
Schwache GML						
Separativ	72 (36.5)	84.5 (5.19)	51 (70.8)	23.9 (5.46)	47.6 (13.3)	53.7 (15.9)
Integrativ	37 (18.8)	84.0 (4.69)	26 (70.3)	24.2 (5.87)	49.5 (12.7)	54.2 (12.4)
Inklusiv	88 (44.7)	82.9 (4.63)	61 (69.3)	25.5 (4.28)	45.1 (13.5)	53.0 (12.3)
Matched Samples						
Separativ	23 (33.3)	84.2 (5.9)	18 (78.3)	33.7 (11.9)	46.2 (12.7)	42.6 (14.7)
Integrativ	23 (33.3)	84.2 (4.7)	18 (78.3)	33.6 (11.5)	50.1 (12.7)	57.0 (10.5)
Inklusiv	23 (33.3)	84.2 (5.6)	18 (78.3)	33.7 (11.2)	44.6 (12.9)	49.9 (11.7)

Anmerkung. Schwache GML = Kinder mit schwacher Graфомotorikleistung zum Prä-Test (Baseline-Messung T1). Matched Samples = Gruppen gematched zu den SchülerInnen mit Graфомotoriktherapie im separativen Setting. Beim Alter handelt es sich um Angaben in Monaten. M = Mittelwert, SD = Standardabweichung.

gnifikanten Verbesserung des Modells ($\chi^2 = 2.78$, $df = 4$, $p = .60$). Aus den Daten zeigt sich deshalb keine ausreichende Evidenz dafür, dass SchülerInnen ohne ausgewiesenen grafomotorischen Förderbedarf im inklusiven oder integrativen Setting durch die Anwesenheit von PsychomotoriktherapeutInnen in der Grafomotorikleistung zusätzlich profitieren (vgl. Tabelle 2).

Settingeffekte auf die Entwicklung von SchülerInnen mit schwacher Grafomotorikleistung

Die Modellberechnungen für die SchülerInnen mit schwacher Grafomotorikleistung zeigten teilweise Probleme mit der Schätzung der zufälligen Effekte (Nullmodell sowie Cross-Level Interaktion), möglicherweise bedingt durch die Homogenisierung der Gruppen (Nullmodell) und den geringeren Stichprobenumfang (Cross-Level Interaktionsmodell), weshalb auch Analysen mit weniger komplexen und robusten Modellen berechnet wurden. Bei all diesen Analysen zeigten sich keine signifikanten Unterschiede zwischen den Settings in Bezug auf die Leistungsentwicklung (vgl. Tabelle 3).

Entwicklung der SchülerInnen mit Grafomotoriktherapie im separativen Setting

Für Fragestellung 3 wurden mittels Propensity Score Matching in Bezug auf die relevanten Merkmale Alter, Geschlecht und grafomotorische Leistung (Baseline) Vergleichsgruppen zu den SchülerInnen mit ausgewiesenem grafomotorischem Förderbedarf im separativen Setting gebildet. Die Ergebnisse zeigen die Vergleichbarkeit der Gruppen hinsichtlich der genannten Merkmale zum Prä-Test (gepoolte Unterschiede in allen Variablen nach dem Matching $p > .99$) sowie die Leistungen zu Post-Test und Follow-up (vgl. Tabelle 1).

Die aus den Nullmodellen geschätzte durchschnittliche Grafomotorikleistung über alle Messzeitpunkte betrug 43.6 Punk-

te (vgl. Tabelle 4), wobei 19% der Varianz innerhalb der SchülerInnen erklärt wurde.

Im anschliessenden Modell mit integriertem Zeitfaktor und Setting entspricht das Intercept der Grafomotorikleistung der SchülerInnen mit ausgewiesenem grafomotorischem Förderbedarf im separativen Setting zum Prä-Test, die erklärte Varianz beträgt 55%. Es zeigen sich unterschiedliche Entwicklungen über die Settings, indem SchülerInnen im integrativen Setting im Follow-up-Test im Vergleich zu den separativ geförderten SchülerInnen einen signifikant stärkeren Leistungszuwachs aufweisen (vgl. Abbildung 2).

Diskussion

Die Ergebnisse zur grafomotorischen Leistungsentwicklung aller SchülerInnen zeigen, dass in allen Settings bedeutsame Fortschritte erzielt werden. Aus theoretischer Sicht ist dieses Ergebnis zu erwarten, weil sich die Grundlagen für die Förderung an den bekannten Gelingensfaktoren für eine erfolgreiche Intervention ausgerichtet haben. Alle Lehrkräfte und PsychomotoriktherapeutInnen erhielten dieselben Fördermaterialien, die zum einen multisensorische Ansätze beinhalten und zum andern kognitive Strategien integrieren. Interventionen, die eine Verknüpfung von sensomotorischen und kognitiven Aufgaben beinhalten, gelten in der Literatur als besonders effektiv (z. B. Diaz Meyer et al., 2017). Entsprechend erstaunt es auch nicht, dass zwischen den Settings keine bedeutsamen Unterschiede zu beobachten sind, weder unmittelbar nach der Interventionsphase noch 6 Monate später in der Follow-up Messung. Vergleichbar fallen die Ergebnisse für die Fragestellung 2 aus. Auch bei Kindern mit schwachen grafomotorischen Leistungen zeigen sich in allen Fördersettings signifikante Fortschritte. Allerdings lassen sich keine signifikanten Unterschiede zwischen den einzelnen Settings feststellen. Mit Blick auf die vorhandenen Untersu-

Tabelle 2
Entwicklung Grafomotorikleistung über die 3 Settings

Level und Variablen		Modell			
	Nullmodell	Random Fixed S	Random Random S	Cross-Level Interaktion	
Level 1					
Intercept	51.3***	[50.4-52.3]			
Post-Test		43.2*** 9.44***	[41.4;45.0] [8.63;10.3]	42.8*** 10.1*** [38.6;45.8] [7.74;16.0]	
Follow-up		12.9***	[12.1;13.7]	13.4*** [8.46;17.8]	
Level 2					
Alter		0.14*	[0.03;0.25]	0.15** [0.04;0.26]	
Geschlecht		-6.93***	[-7.9; -5.9]	-6.98*** [-8.0;-6.0]	
Level 3					
Integratives Setting		0.28	[-2.4;2.9]	1.07 [-4.2;6.3]	
Inklusives Setting		1.31	[-0.9;3.5]	2.03 [-2.5;6.5]	
Integratives Setting x Post-Test				-2.16 [-8.2;3.9]	
Integratives Setting x Follow-up				-0.24 [-7.1;6.6]	
Inklusives Setting x Post-Test				-2.67 [-7.8;2.5]	
Inklusives Setting x Follow-up				0.57 [-5.3;6.4]	
Zufällige Effekte					
σ ²	150.7	107.1	64.0	64.0	
T ₀₀ KlassenID.SchülerInnenID	43.6	45.8	59.8	59.8	
T ₀₀ KlassenID	14.1	13.5	69.1	70.4	
T ₁₁ KlassenID.Post-Test			94.9	96.0	
T ₁₁ KlassenID.Follow-up			122.5	125.6	
ρ ₀₁ KlassenID Post-Test			-0.82	-0.83	
ρ ₀₁ KlassenID Follow-up			-0.86	-0.87	
ICC	0.28	0.36	0.62	0.62	
N KlassenID	87	87	87	87	
N SchülerInnenID	1332	1332	1332	1332	
-2 log likelihood	30610	29574***	28686***	28684	
Beobachtungen	3784	3784	3784	3784	
Marginales R ² / Konditionales R ²	.000 / .277	.201 / .486	.210 / .699	.211 / .700	

Anmerkung. I = Intercept; S = Slope. 95%-Konfidenzintervall in Klammern. * = $p < .05$; ** = $p < .01$; *** = $p < .001$.

Tabelle 3
Entwicklung Grafomotorikleistung über die 3 Settings: Unterdurchschnittliche Gruppe

Level und Variablen		Modell			
		Nullmodell	Random I Fixed S	Random I Random S	Cross-Level Interaktion
Level 1					
Intercept		40.9***	24.2***	24.5***	24.3***
Post-Test			22.1***	19.4***	20.9***
Follow-up			28.8***	26.9***	27.1***
Level 2					
Alter			-0.29*	-0.29**	-0.30**
Geschlecht			-5.33***	-4.75***	-4.77***
Level 3					
Integratives Setting			1.19	0.20	0.04
Inklusives Setting			-1.20	1.33	0.99
Integratives Setting x Post-Test					2.34
Integratives Setting x Follow-up					3.36
Inklusives Setting x Post-Test					-3.72
Inklusives Setting x Follow-up					-1.71
Zufällige Effekte					
σ^2		272.9	87.5	56.4	56.5
T ₀₀ KlassenID:SchülerInnenID		0.00	15.4	20.4	20.3
T ₀₀ KlassenID		9.03	18.5	0.31	0.00
T ₁₁ KlassenID:Post-Test				79.3	72.9
T ₁₁ KlassenID:Follow-up				82.8	86.8
P ₀₁ KlassenID: Post-Test					N/A
P ₀₁ KlassenID: Follow-up					N/A
ICC		N/A	0.28	0.56	N/A
N _{KlassenID}		58	58	58	58
N _{SchülerInnenID}		197	197	197	197
-2 log likelihood		4837	4292***	4197***	4194
Beobachtungen		571	571	571	571
Marginales R ² / Konditionales R ²		.000 / N/A	.573 / .692	.517 / .788	.714 / N/A

Anmerkung. I = Intercept; S = Slope. N/A = Keine Berechnung möglich. 95%-Konfidenzintervall in Klammern. * = $p < .05$; ** = $p < .01$; *** = $p < .001$.

Tabelle 4
SchülerInnen mit ausgewiesenem grafomotorischem Förderbedarf im separativen Setting vs. Matched Samples aus integrativen / inklusiven Settings

Level und Variablen	Nullmodell		Random I Fixed S	
Level 1				
Intercept	43.6***	[41.1;46.0]	33.7***	[28.4;38.9]
Post-Test			12.2***	[6.24;18.3]
Follow-up			8.91**	[2.91;7.44]
Level 2				
Integratives Setting			0.01	[-7.4;7.4]
Inklusives Setting			0.01	[-7.4;7.4]
Integratives Setting x Post-Test			6.97	[-1.7;15.6]
Integratives Setting x Follow-up			12.0 **	[3.33;20.6]
Inklusives Setting x Post-Test			-1.48	[-10.2;7.2]
Inklusives Setting x Follow-up			7.78	[-1.0;16.5]
Zufällige Effekte				
σ ²	180.4		102.2	
T ₀₀ SchülerInnenID	41.1		62.8	
ICC	0.19		0.38	
N SchülerInnenID	69		69	
-2 log likelihood	1695		1575***	
Beobachtungen	207		207	
Marginales R ² / Konditionales R ²	.000 / .185		.275 / .551	

Anmerkung. I = Intercept; S = Slope. 95%-Konfidenzintervall in Klammern. * = $p < .05$; ** = $p < .01$; *** = $p < .001$.

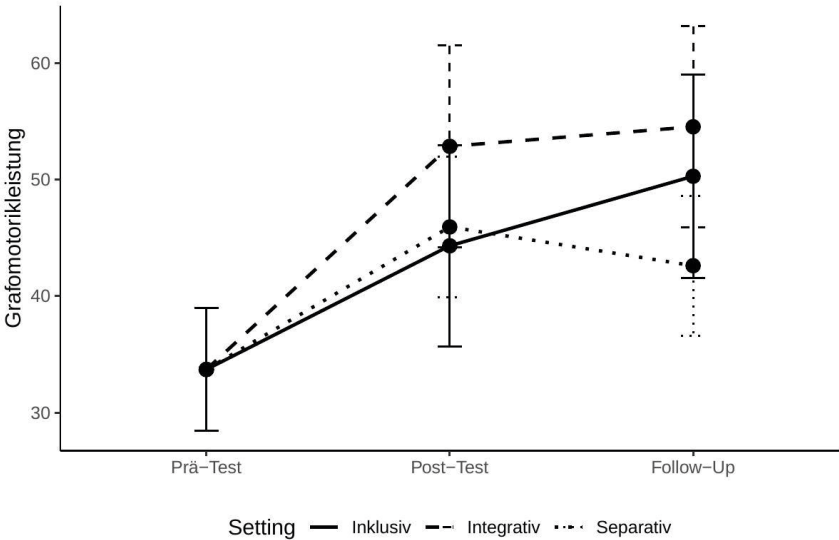


Abbildung 2
Grafomotorische Leistungsentwicklung von SchülerInnen im separativen Fördersetting im Vergleich zu gematchten SchülerInnen in integrativen/inklusive Fördersettings

chungsergebnisse und deren theoretische Einbettung hätte vermutet werden können, dass bezogen auf diese Kinder, integrative und inklusive Settings vorteilhaft sind. Ein solcher Vorteil ist nicht zu beobachten, alle Kinder profitieren von der Förderung.

Von besonderem Interesse sind die Ergebnisse zu Fragestellung 3. Hier wird nun der Fokus auf die Kinder mit Therapie in separativen Settings gelegt. Diese Kinder besuchen klassenextern ein therapeutisches Angebot. Aus theoretischer Sicht hätte angenommen werden können, dass die Kinder mit einem ausgewiesenen Förderbedarf von diesem spezifischen Therapieangebot mehr profitieren als vergleichbare Kinder in integrativen oder inklusiven Settings. Die Ergebnisse zeigen, dass sich die grafomotorische Leistung der SchülerInnen in den verschiedenen Fördersettings unmittelbar nach der Intervention ähnlich entwickeln. Bei der Betrachtung der Daten nach sechs Monaten (Follow-up) lassen sich jedoch statistisch bedeutsame Unterschiede zwischen den verschiedenen Fördersettings feststellen. Dabei profitieren die integriert geförderten SchülerInnen am meisten, gefolgt von den SchülerInnen in inklusiven und separativen Fördersettings. Dieses Ergebnis lässt sich möglicherweise durch den sogenannten Fade-out-Effekt erklären, der als ein Muster abnehmender Interventionseffekte nach dem Ende einer Intervention definiert wird (Elango et al., 2015). Ähnlich wie bei der vorliegenden Untersuchung zeigen viele Studien, die kompetenzorientierte Interventionen untersuchen, dass die Effekte am Ende der Intervention abnehmen (Bailey et al., 2020). Eine inhaltliche Begründung für diesen Effekt kann in der Fokussierung auf andere Fähigkeiten nach dem Ende der Intervention abgeleitet werden. Dadurch verlieren die spezifisch trainierten Fähigkeiten an Bedeutung, was eine Abnahme zur Folge hat (Wixted, 2004). Unsere Ergebnisse zeigen eine solche Abnahme beim separativen Fördersetting und angedeutet ebenfalls beim integrativen Fördersetting. Aufgrund der vorhandenen Unterrichtsprotokolle kann

angenommen werden, dass in allen Fördersettings während der Interventionsphase intensiv mit den eingeführten Fördermaterialien gearbeitet wurde. Nach Abschluss der Intervention wurde in vielen Klassen die Förderung wieder in der zuvor praktizierten Form umgesetzt. Das heisst, dass im separativen und wahrscheinlich auch in einigen integrativen Fördersettings wieder die herkömmliche therapeutische Förderung erfolgte, bei der die Grafomotorik lediglich ein Teilaspekt unter mehreren Förderschwerpunkten ist (z. B. Körperwahrnehmung, Grobmotorik, sozio-emotionale Kompetenzen). Es kann vermutet werden, dass bei einer konsequenten Fortführung der Intervention in separativen Settings ein Absinken der Leistungskurve hätte verhindert werden können. Im Gegensatz dazu wurden im inklusiven Fördersetting nicht nur die Fördermaterialien angepasst, sondern es wurde auch ein spezifisches didaktisches Unterrichtsarrangement entwickelt. So wurde in den inklusiven Fördersettings an und mit einer gemeinsamen Lernumgebung gelernt. Es kann davon ausgegangen werden, dass das inklusive Fördersetting in vielen Klassen auch nach der Intervention fortgeführt wurde, was den kontinuierlichen Anstieg der grafomotorischen Leistungen bis zum Follow-up erklären könnte. Zusammenfassend legen die analysierten Daten den Schluss nahe, dass eine separative therapeutische Förderung von Kindern mit ausgewiesenem grafomotorischem Förderbedarf zu keinen Vorteilen führt, wenn mit der grafomotorischen Leistungsentwicklung in integrativen oder inklusiven Settings verglichen wird. Vielmehr muss angenommen werden, dass längerfristig eine integrative oder inklusive Förderung mindestens zu gleichen Fortschritten in der Grafomotorik führen kann.

Limitationen

Die vorliegenden Auswertungen konzentrieren sich auf die Wirkung unterschiedlicher Fördersettings in Bezug auf grafomotorische Leistungen. Dieser Fokus ist begrenzt

und zeigt damit nur einen Ausschnitt. Zukünftige Studien sollten daher zusätzliche abhängige Variablen, wie etwa das Selbstkonzept, berücksichtigen, um ein umfassenderes Bild zu erhalten. Weiter konnte die Zuordnung zu den Settings nur teilweise zufällig erfolgen. Bei der Zuweisung mussten die Gegebenheiten vor Ort in den Schulen berücksichtigt werden. Lehrkräfte wurden nicht gezwungen, sich einem Setting zuzuordnen, dies hätte möglicherweise negative Auswirkungen auf die Motivation gehabt. Ein potenzieller Einflussfaktor der vorliegenden Ergebnisse ist die Neuartigkeit der inklusiven und integrativen Fördersettings, die speziell für die Untersuchung konzipiert wurden, während das separate Fördersetting der herkömmlichen, etablierten Förderpraxis entsprach. Zukünftige Längsschnittstudien sollten daher prüfen, ob sich die Effekte langfristig verändern, sobald die Fördersettings nicht mehr als neu angesehen werden.

Schlussfolgerungen und Ausblick

Die Ergebnisse der vorliegenden Untersuchung tragen zur aktuellen Integrations- und Inklusionsdiskussion bei, indem sie die Wirksamkeit verschiedener Fördersettings im Bereich der Grafomotorik beleuchten. Während die traditionelle separate bzw. therapeutische Förderung seit den Nachkriegsjahren in der Schweiz etabliert ist und darauf abzielt, Kinder mit spezifischem Förderbedarf gezielt zu unterstützen, gewinnen integrative und inklusive Ansätze zunehmend an Bedeutung. Bislang fehlen jedoch empirische Befunde zur Wirksamkeit dieser verschiedenen Fördersettings. Der Vergleich separativer, integrativer und inklusiver grafomotorischer Fördersettings anhand unterschiedlicher Stichproben ermöglicht ein erweitertes Verständnis der Auswirkungen dieser Fördersettings auf die grafomotorische Leistung. Ob separate Settings auch andere Aspekte, wie das Selbstkonzept, beeinflussen, bleibt jedoch unklar und wird in weiteren Analysen vertieft untersucht.

Literatur

- Achymy, Z. I., Kadar, M., Razaob, N. A., & Wan Yunus, F. (2022). Factors influencing handwriting development among preschool children: A systematic review. *Kesmas: Jurnal Kesehatan Masyarakat Nasional*, 17(4), 235. <https://doi.org/10.21109/kesmas.v17i4.6209>
- Aguinis, H., Gottfredson, R.K. and Culpepper, S.A. (2013) Best-practice recommendations for estimating cross-level interaction effects using multilevel modeling. *Journal of Management*, 39, 1490–1528. <https://doi.org/10.1177/0149206313478188>
- Bailey, D. H., Duncan, G. J., Cunha, F., Foorman, B. R., & Yeager, D. S. (2020). Persistence and fade-out of educational-intervention effects: Mechanisms and potential solutions. *Psychological Science in the Public Interest*, 21(2), 55–97. <https://doi.org/10.1177/1529100620915848>
- Becker, D. R., Miao, A., Duncan, R., & McClelland, M. M. (2014). Behavioral self-regulation and executive function both predict visuomotor skills and early academic achievement. *Early Childhood Research Quarterly*, 29(4), 411–424. <https://doi.org/10.1016/j.jecresq.2014.04.014>
- Beery, K. E., Buktenica, N. A., & Beery, N. A. (2010). *Beery-Buktenica developmental test of visual-motor integration (Beery VMI)* (6th ed.). Pearson.
- Behrens, U. (2017). Entwicklung der Schreibkompetenz: Vorschule und Primarstufe. In M. Becker-Mrotzek, J. Grabowski, & T. Steinhoff (Hrsg.), *Forschungshandbuch empirische Schreibdidaktik* (S. 75–89). Waxmann.
- Berninger, V. W. (Hrsg.). (2006). *Past, present and future contributions of cognitive writing research to cognitive Psychology*. Psychology Press.
- Berninger, V. W., Nagy, W., Tanimoto, S., Thompson, R., & Abbott, R. D. (2015). Computer instruction in handwriting, spelling, and composing for students with specific learning disabilities in grades 4–9. *Computers & Education*, 81, 154–168. <https://doi.org/10.1016/j.compedu.2014.10.005>
- Berninger, V. W., & Winn, W. D. (2006). Implications and advancements in brain research and technology for writing development, writing instruction and educational evolution. In S. Graham & J. Fitzgerald (Hrsg.), *Handbook of Writing Research* (S. 96–114). Guilford Press.

- Bless, G. (2018). Wirkungen der schulischen Integration auf Schülerinnen und Schüler. *Schweizerische Zeitschrift für Heilpädagogik*, 24(2).
- Booth, T., & Ainscow, M. (2019). *Index für Inklusion: Ein Leitfaden für Schulentwicklung* (2. Auflage). Beltz.
- Case-Smith, J., Holland, T., Lane, A., & White, S. (2012). Effect of a coteaching handwriting program for first graders: one-group pretest–posttest design. *The American Journal of Occupational Therapy*, 66(4), 396–405. <https://doi.org/10.5014/ajot.2012.004333>
- Case-Smith, J., Holland, T., & White, S. (2014). Effectiveness of a co-taught handwriting program for first grade students. *Physical & Occupational Therapy In Pediatrics*, 34(1), 30–43. <https://doi.org/10.3109/01942638.2013.783898>
- Chandler, M. C., Gerde, H. K., Bowles, R. P., McRoy, K. Z., Pontifex, M. B., & Bingham, G. E. (2021). Self-regulation moderates the relationship between fine motor skills and writing in early childhood. *Early Childhood Research Quarterly*, 57, 239–250. <https://doi.org/10.1016/j.ecresq.2021.06.010>
- Dalgaard, N. T., Bondebjerg, A., Viinholt, B. C. A., & Filges, T. (2022). The effects of inclusion on academic achievement, socioemotional development and wellbeing of children with special educational needs. *Campbell Systematic Reviews*, 18(4), 1–44. <https://doi.org/10.1002/cl2.1291>
- Diaz Meyer, M., Schneider, M., Marquardt, C., Knopf, J., & Luptowicz, C. (2017). Schreibmotorische Förderung bei Erstklässlern: Ergebnisse einer Interventionsstudie. *Didaktik Deutsch*, 22(43), 33–56.
- Dinehart, L. H. (2014). Handwriting in early childhood education: Current research and future implications. *Sage Journals*, 15(1), 97–118.
- Duiser, I. H. F., Van Der Kamp, J., Ledebt, A., & Savelsbergh, G. J. P. (2014). Relationship between the quality of children's handwriting and the Beery Buktenica developmental test of visuomotor integration after one year of writing tuition. *Australian Occupational Therapy Journal*, 61(2), 76–82. <https://doi.org/10.1111/1440-1630.12064>
- Eckhart, M., & Sägeser, J. (2016). Förderplanung im Unterricht. *Schweizerische Zeitschrift für Heilpädagogik*, 22(2), 13–19.
- Elango, S., García, J. L., Heckman, J. J., & Hojman, A. (2015). Early childhood education. IZA Discussion Paper No. 9476, Verfügbar unter: <https://ssrn.com/abstract=2690725>
- Haeberlin, U. (2005). *Grundlagen der Heilpädagogik. Einführung in eine wertgeleitete erziehungswissenschaftliche Disziplin*. Haupt.
- Krämer, S., Möller, J., & Zimmermann, F. (2021). Inclusive education of students with general learning difficulties: A meta-analysis. *Review of Educational Research*, 91(3), 432–478. <https://doi.org/10.3102/0034654321998072>
- Kranz, I., Sammann, K., Amft, S., & Vetter, M. (2011). Effekte einer grafomotorisch ausgerichteten psychomotorischen Intervention bei Kindern im Vorschulalter. *Psychologie in Erziehung und Unterricht*, 58(2), 139–151. <https://doi.org/10.2378/peu2011.art05d>
- Kushki, A., Schwellnus, H., Ilyas, F., & Chau, T. (2011). Changes in kinetics and kinematics of handwriting during a prolonged writing task in children with and without dysgraphia. *Research in Developmental Disabilities*, 32(3), 1058–1064. <https://doi.org/10.1016/j.ridd.2011.01.026>
- Landis, J. R. & Koch, G. G. (1977) An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 33, 363–374. <https://doi.org/10.2307/2529786>
- Maurer, M. N., Truxius, L., Sägeser Wyss, J., & Eckhart, M. (2023). From scribbles to script: graphomotor skills' impact on spelling in early primary school. *Children*, 10(12), 1886. <https://doi.org/10.3390/children10121886>
- McCutchen, D. (2011). From novice to expert: Language and memory processes in the development of writing skill. *Journal of Writing Research*, 3(1), 51–68. <https://doi.org/10.17239/jowr-2011.03.01.3>
- Sägeser Wyss, J., Eckhart, M., & Maurer, M. N. (2024). *Screening und Differentialdiagnostik der Grafomotorik im schulischen Kontext—2. Instrument zur Erfassung des grafomotorischen Entwicklungsstandes bei Kindern zwischen 4-9 Jahren* (1. Auflage). Hogrefe.
- Sägeser Wyss, J., Lozano, C. S., & Simovic, L. (2021a). *GRAFINK - Grafomotorik und Inklusion. Grundlagen und Materialien für das Erstschreiben*. Hogrefe.
- Sägeser Wyss, J., Sahli Lozano, C. & Eckhart, M.

- (2021b). Grafomotorik und schulische Inklusion. Entwicklung und Anwendung eines Rahmenmodells. In M. Vetter, B. Wuttig, L. Spahn, & H. Göhle (Hrsg.), *Profession – Gender – Inklusion. Motologie und Psychomotorik – Marburger Beiträge zu einer responsiven Fachentwicklung. Forschung und Theoriebildung*. Band 1.
- Sahli Lozano, C., Gruber, L., & Wüthrich, S. (2017). Subjektiv wahrgenommenes Integriertsein von Kindern in Schulsystemen mit integrativen Massnahmen. *Empirische Pädagogik*, 31(3), 284–302.
- Salas, N., & Silvente, S. (2020). The role of executive functions and transcription skills in writing: A cross-sectional study across 7 years of schooling. *Reading and Writing*, 33(4), 877–905. <https://doi.org/10.1007/s11145-019-09979-y>
- Skar, G. B., Lei, P.-W., Graham, S., Aasen, A. J., Johansen, M. B., & Kvistad, A. H. (2022). Handwriting fluency and the quality of primary grade students' writing. *Reading and Writing*, 35(2), 509–538. <https://doi.org/10.1007/s11145-021-10185-y>
- Sturm, A., Nänni, R., & Wyss, S. (2017). Förderung hierarchieniedriger Schreibprozesse. In M. Philipp (Hrsg.), *Handbuch Schriftspracherwerb und weiterführendes Lesen und Schreiben* (S. 266–284). Beltz Juventa.
- Taverna, L., Tremolada, M., Dozza, L., Zanin Scaratti, R., Ulrike, D., Lallo, C., & Tosetto, B. (2020). Who benefits from an intervention program on foundational skills for handwriting addressed to kindergarten children and first graders? *International Journal of Environmental Research and Public Health*, 17(6), 1–24. <https://doi.org/10.3390/ijerph17062166>
- UNO (2006). Übereinkommen über die Rechte von Menschen mit Behinderungen. <https://www.adm-in.ch/opc/de/classified-compilation/20122488/index.html#>
- Van Galen, G. P. (1991). Handwriting: Issues for a psychomotor theory. *Human Movement Science*, 10(2–3), 165–191. [https://doi.org/10.1016/0167-9457\(91\)90003-G](https://doi.org/10.1016/0167-9457(91)90003-G)
- Vetter, M., Amft, S., Sammann, K., & Kranz, I. (2021). *G-Fipps: Grafomotorische Förderung – Ein psychomotorisches Praxisbuch* (3. Auflage).
- Vetter, M., Kranz, I., Sammann, K., Amft, S., Hättich, A., & Venetz, M. (2009). *Zur Wirksamkeit grafomotorischer Förderung in integrativ und präventiv ausgerichteter Psychomotorik*.
- Wixted, J. T. (2004). The psychology and neuroscience of forgetting. *Annual Review of Psychology*, 55, 235–269.

Autorinnen- und Autorenhinweis

Der vorliegende Artikel entstand im Rahmen des vom Schweizerischen Nationalfonds finanzierten Forschungsprojekts Settings der Förderung der Grafomotorik (grafset) an der Pädagogischen Hochschule Bern (Laufzeit: 2020 – 2024, Nr. 100019_189187). Die Autoren erklären, dass kein Interessenkonflikt vorliegt.

 Eckhart Michael
<https://orcid.org/0009-0000-8672-2984>

 Joséphine Schwery
<https://orcid.org/0009-0004-3859-8017>

 Judith Sägger Wyss
<https://orcid.org/0009-0002-6732-7916>

 Sergej Wüthrich
<https://orcid.org/0000-0002-3301-8587>

Korrespondenzadresse

Michael Eckhart

Institutsleiter, Pädagogische Hochschule Bern
Fabrikstrasse 8
CH-3012 Bern
michael.eckhart@phbern.ch

Erstmals eingereicht: 17.12.2024
Überarbeitung eingereicht: 20.08.2025
Angenommen: 06.08.2025

Offene Daten	Die Daten stehen nach endgültigem Projektabschluss auf SwissUBase zur Verfügung.
Offener Code	Die Computercodes sind auf Anfrage verfügbar (josephine.schwery@phbern.ch).
Offene Materialien	Die Materialien sind auf Anfrage verfügbar (judith.saegesser@phbern.ch).
Präregistrierung	Die Studie wurde nicht präregistriert.
Votum Ethikkommission	Die Studie wurde von der Ethikkommission der Pädagogischen Hochschule Bern genehmigt (Referenznummer: 19s000201).
Finanzielle und weitere sachliche Unterstützung	Die Studie wurde vom Schweizerischen Nationalfonds (SNF) finanziert (Laufzeit 2020-2023, Projekt Nr. 100019_189187)
Autorenschaft	Michael Eckhart: Entwurf Forschungsdesign, Durchführung Analysen, Verfassen und Überarbeitung des Manuskripts, Supervision, organisatorische und administrative Steuerung des Projekts, Einwerbung von Fördermitteln; Joséphine Schwery Klingele: Datenaufbereitung, Durchführung Analysen, Durchführung der Datenerhebung, Erstellung von Abbildungen, Verfassen und Überarbeitung des Manuskripts; Judith Säegger Wyss: Durchführung der Datenerhebung, Entwicklung von Messinstrumenten, Verfassen und Überarbeitung des Manuskripts, organisatorische und administrative Steuerung des Projekts, Einwerbung von Fördermitteln; Sergej Wüthrich: Entwurf Forschungsdesign, Durchführung Analysen, Erstellung Grafiken und Tabellen, Verfassen und Überarbeitung des Manuskripts
KI und generative Modelle	Es wurden keine generativen Modelle für die Analysen genutzt.