

Wilbert, Jürgen; Börnert-Ringleb, Moritz; Lüke, Timo
**The power of randomization tests to detect individual intervention effects.
Insights from a Monte Carlo study**
Empirische Sonderpädagogik 17 (2025) 3, S. 184-199



Quellenangabe/ Reference:

Wilbert, Jürgen; Börnert-Ringleb, Moritz; Lüke, Timo: The power of randomization tests to detect individual intervention effects. Insights from a Monte Carlo study - In: Empirische Sonderpädagogik 17 (2025) 3, S. 184-199 - URN: urn:nbn:de:0111-pedocs-353109 - DOI: 10.25656/01:35310; 10.2440/003-0048

<https://nbn-resolving.org/urn:nbn:de:0111-pedocs-353109>

<https://doi.org/10.25656/01:35310>

in Kooperation mit / in cooperation with:



<https://www.psychologie-aktuell.com/journale/empirische-sonderpaedagogik.html>

Nutzungsbedingungen

Dieses Dokument steht unter folgender Creative Commons-Lizenz: <http://creativecommons.org/licenses/by-nc/4.0/deed.de> - Sie dürfen das Werk bzw. den Inhalt vervielfältigen, verbreiten und öffentlich zugänglich machen sowie Abwandlungen und Bearbeitungen des Werkes bzw. Inhaltes anfertigen, solange Sie den Namen des Autors/Rechteinhabers in der von ihm festgelegten Weise nennen und das Werk bzw. den Inhalt nicht für kommerzielle Zwecke verwenden.

Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use

This document is published under following Creative Commons-License: <http://creativecommons.org/licenses/by-nc/4.0/deed.en> - You may copy, distribute and render this document accessible, make adaptations of this work or its contents accessible to the public as long as you attribute the work in the manner specified by the author or licensor. You are not allowed to make commercial use of the work, provided that the work or its contents are not used for commercial purposes.

By using this particular document, you accept the above-stated conditions of use.



Kontakt / Contact:

peDOCS
DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation
Informationszentrum (IZ) Bildung
E-Mail: pedocs@dipf.de
Internet: www.pedocs.de

Mitglied der


Leibniz-Gemeinschaft

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

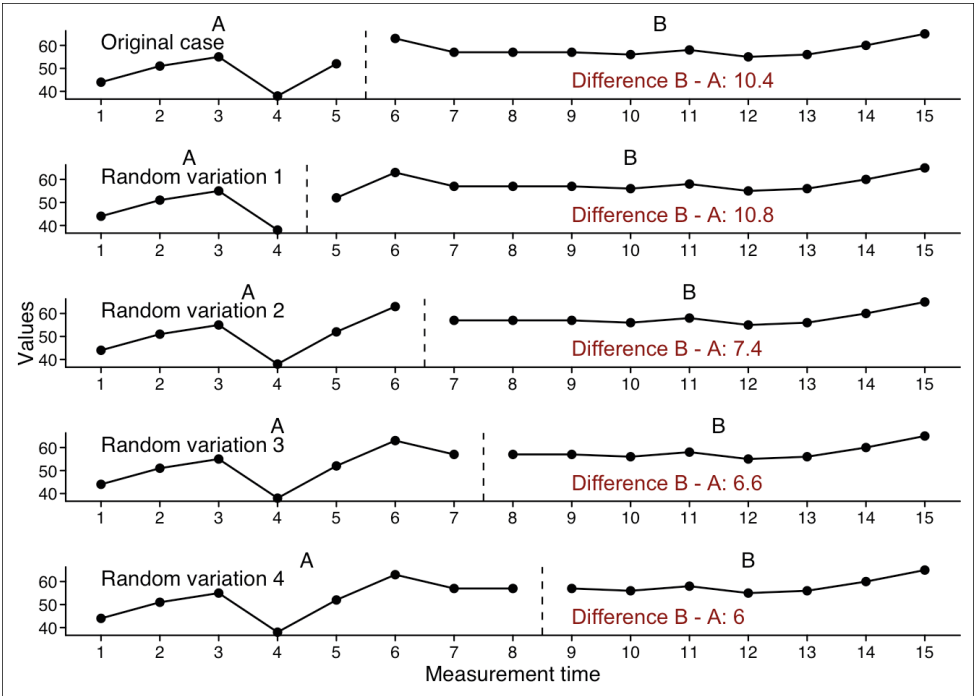


Figure 1
Illustration of the randomization of a single-case study and the resulting distribution of mean differences

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.

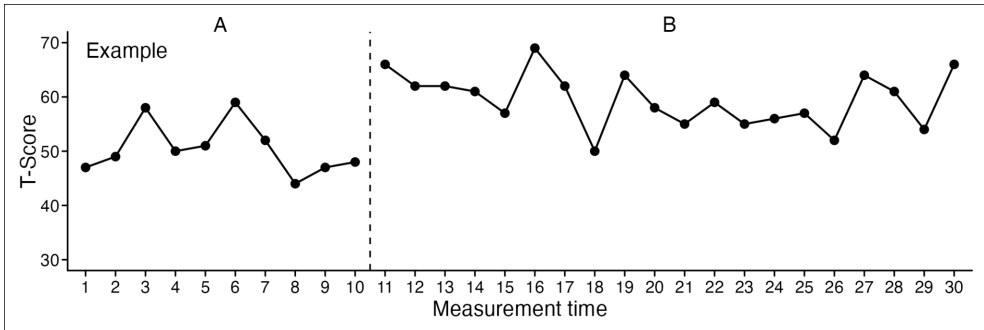


Figure 2

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 ($d =$

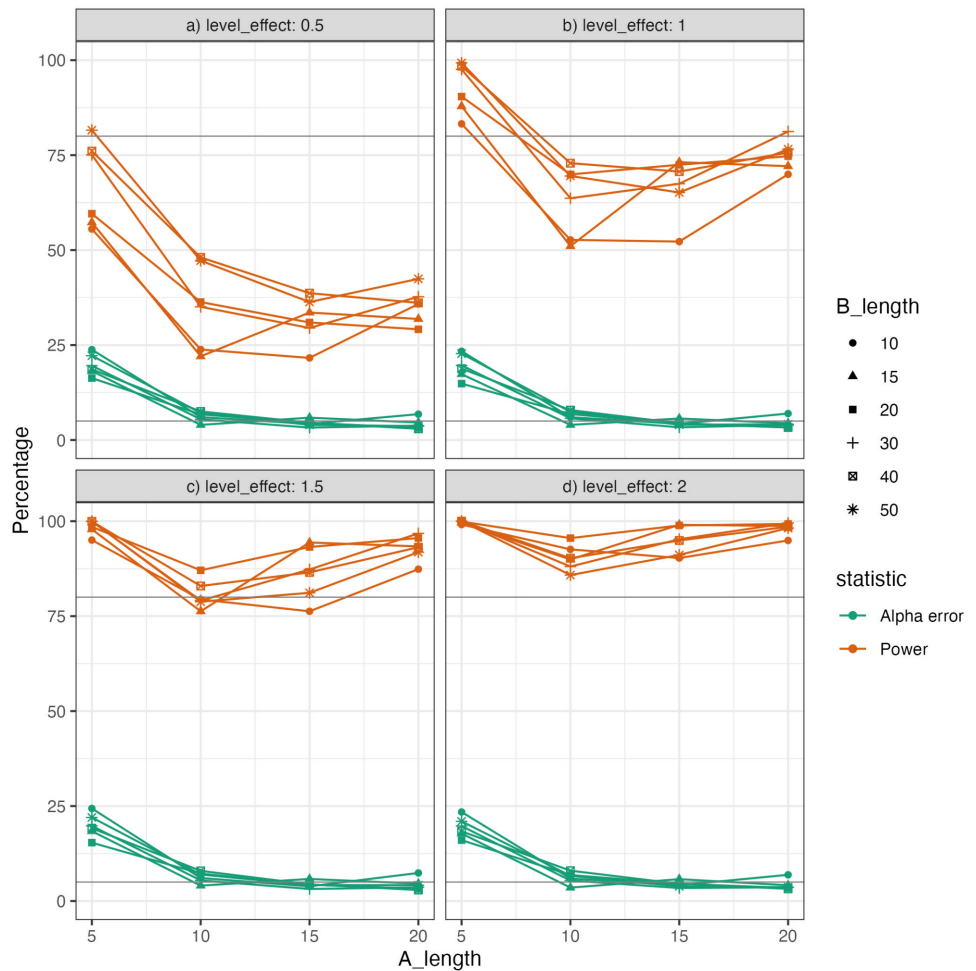


Figure 3
Power and alpha error rates (line color) for different intervention effect sizes (panel) and measurement times per phase (dot shape and x-axis; simulation 1)

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 ($r_{tt} =$

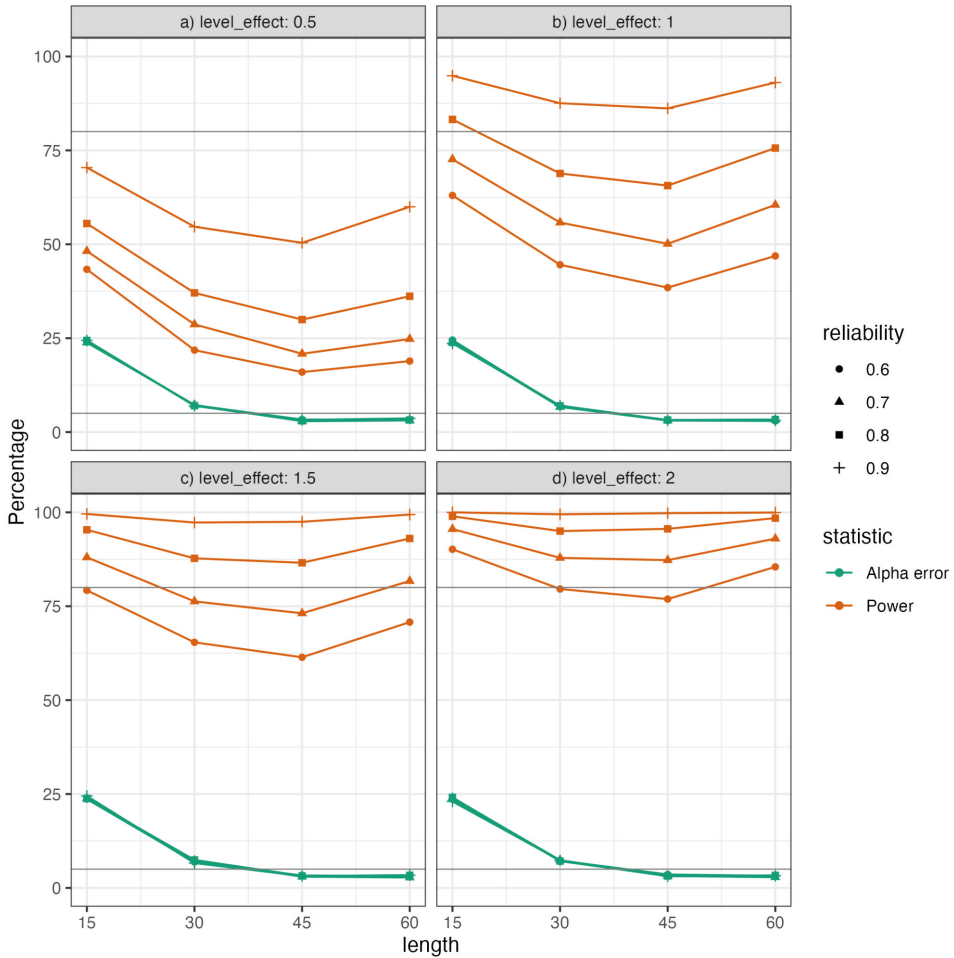


Figure 4
Power and alpha error rates (line color) for different intervention effect sizes (part), measurement times (x-axis), and measurement reliabilities (dot shape; simulation 2)

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

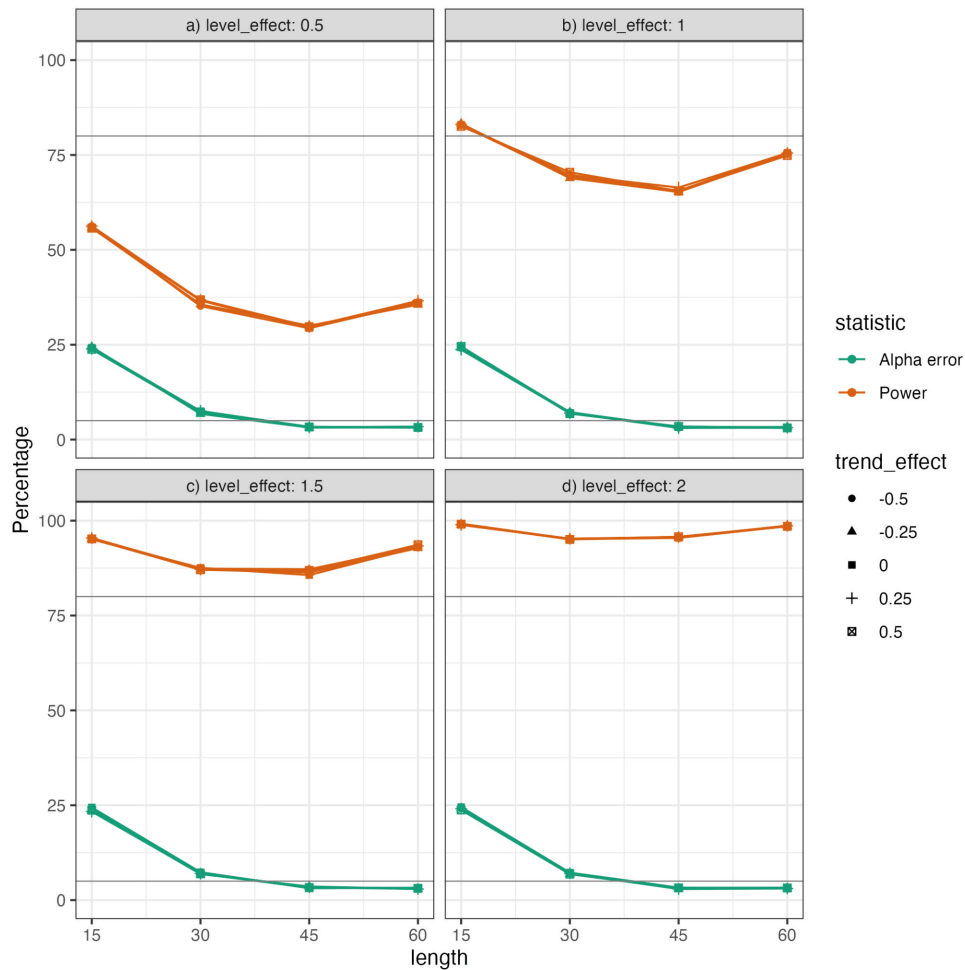


Figure 5
Power and alpha error rates (line color) for different intervention effect sizes (part), measurement times (x-axis), and different trend effect sizes (dot shape; simulation 3)

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

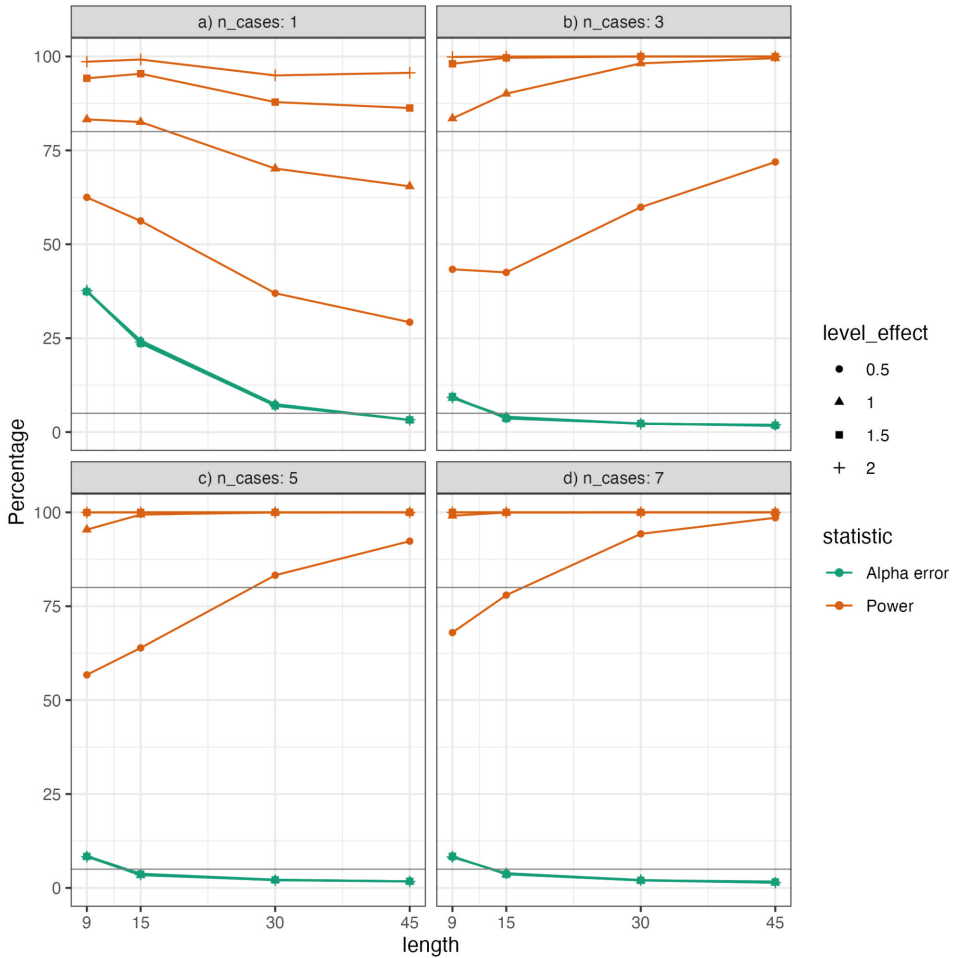


Figure 6
Power and alpha error rates (line color) for different numbers of cases per study (part), measurement times (x-axis), and various intervention effect sizes (dot shape; simulation 4)

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.