J Appl Physiol 139: 1220–1227, 2025.
First published September 30, 2025; doi:10.1152/japplphysiol.00485.2025

american physiological society

JOURNAL OF APPLIED PHYSIOLOGY

REVIEW

Concerns regarding the standard deviation of individual responses for assessing treatment response heterogeneity

Aaron R. Caldwell,[1] David B. Allison,[2] Andrew W. Brown,[3,4] Gary L. Gadbury,[5] Thirupathi Reddy Mokalla,[2] R. Drew Sayer,[6] and Andrew D. Vigotsky[7]

[1]*Department of Biostatistics, Fay W. Boozman College of Public Health, University of Arkansas for Medical Sciences Northwest, Springdale, Arkansas, United States;* [2]*Department of Epidemiology and Biostatistics, School of Public Health-Bloomington, Indiana University, Bloomington, Indiana, United States;* [3]*Department of Biostatistics, University of Arkansas for Medical Sciences, Little Rock, Arkansas, United States;* [4]*Arkansas Children's Research Institute, Little Rock, Arkansas, United States;* [5]*Department of Statistics, Kansas State University, Manhattan, Kansas, United States;* [6]*Department of Family and Community Medicine, Heersink School of Medicine, University of Alabama at Birmingham, Birmingham, Alabama, United States; and* [7]*Neuroscience Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States*

Abstract

The estimation of treatment response heterogeneity (TRH) is increasingly important as medicine moves toward personalized approaches. Although various statistical methods have been proposed to quantify TRH in parallel-group trials, the standard deviation of individual responses (SD_{IR}) has gained prominence within physiological research. This method is intended to quantify individual response variation by comparing standard deviations of change scores between intervention and control groups. We acknowledge that SD_{IR} represents an improvement over many other flawed approaches that often involve responder counting. However, SD_{IR} has critical limitations: *1*) it cannot overcome the fundamental problem of causal inference because the correlation between potential outcomes remains unidentifiable, *2*) it is incorrectly predicated on the assumption that TRH is present only when treatment group variance exceeds control group variance, and *3*) it is statistically inefficient. We present an alternative framework, which involves assessing heteroskedasticity and estimating the bounds for the standard deviation of treatment effects (SD_D). The presence of heteroskedasticity between treatment groups is a sufficient but not necessary condition for the presence of TRH. Furthermore, SD_D makes fewer assumptions than SD_{IR} and, therefore, paints a more complete picture of potential TRH. Using data from a published exercise physiology study, we demonstrate how SD_D can better characterize uncertainty in TRH estimation. We recommend that researchers probe TRH by assessing heteroskedasticity, providing bounds for SD_D, and estimating outcome distributions and probabilities while carefully crafting the theoretical rationale for the presence of TRH.

individual differences; individual response variation; personalized medicine; potential outcomes; responders

INTRODUCTION

Treatment response heterogeneity (TRH) has become increasingly important with the rise of personalized medicine. Although various statistical approaches have been proposed to quantify this heterogeneity in parallel-group trials, one method has gained prominence in exercise physiology research: the standard deviation of individual responses (SD_{IR}). This method, popularized by Atkinson and Batterham (1) and Hopkins (2), is intended to quantify "true" individual response variation by comparing the standard deviations of change scores between intervention and control groups.

The SD_{IR} approach has been used mainly in exercise physiology research, where it has been used to identify so-called "responders" and "nonresponders" to interventions and estimate the proportion of individuals who might benefit from treatment (3, 4). Those using SD_{IR} in situations where there is a negative control (such as a no-treatment or placebo group) implicitly assume that such groups exhibit zero true interindividual outcome variance, and any observed variance can be attributed solely to measurement error and within-subject variability (e.g., biological variation). However, despite its popularity, the SD_{IR} method rests on several other crucial assumptions that can limit its validity and utility.

In this paper, we critically evaluate the statistical and conceptual foundations of the SD_{IR} method. Specifically, we show that: *1*) SD_{IR} has strict assumptions about TRH, *2*) this may lead researchers to dismiss the potential of TRH when it is present, and *3*) the method is inefficient compared with other heteroskedasticity tests (i.e., statistical methods that

Correspondence: A. R. Caldwell (caldwellaaron@uams.edu).
Submitted 2 June 2025 / Revised 14 August 2025 / Accepted 22 September 2025

assess how residual variance depends on an independent variable, such as group membership).

We then present alternative approaches that address these limitations and allow researchers to estimate TRH without such strict assumptions. Based on bounds for the true standard deviation of treatment effects, these methods provide a more complete picture of potential response variation while making fewer assumptions about its underlying nature.

Our goal is to encourage the investigation of TRH using approaches that rest on sound statistical foundations and have transparent and justifiable assumptions. By acknowledging the limitations of current methods and exploring approaches that explicitly account for uncertainty, researchers can enhance their contributions to the investigation of personalized treatment approaches.

DEFINING TREATMENT RESPONSE HETEROGENEITY

In parallel-group trials, understanding TRH requires a precise distinction between true treatment effects (i.e., treatment response) and observed changes. Consider a pre-post parallel-group design comparing treatment (T) and control (C). By the end of the study, an individual has two potential outcomes: $Y(\mathrm{T})$ if assigned to treatment and $Y(\mathrm{C})$ if assigned to control. The true individual treatment effect is the difference between these potential outcomes: $D = Y(\mathrm{T}) - Y(\mathrm{C})$. TRH represents how D varies across individuals. Importantly, the true individual D cannot be directly observed in parallel-group trials because each participant receives only one condition. This is true even in a standard two-condition two-period crossover trial because each participant receives only one condition in one period of time. Within each study participant, only one outcome can be observed in any period of time; the other (potential) outcome is the counterfactual (i.e., what would have happened had the participant been randomized to the other group).

A common misconception is attributing observed within-group changes to the treatment (1). Researchers often consider the pre-post change in the treatment group $[Y(\mathrm{T})_{\mathrm{Post}} - Y(\mathrm{T})_{\mathrm{Pre}}]$ as representing D and the change in the control group $[Y(\mathrm{C})_{\mathrm{Post}} - Y(\mathrm{C})_{\mathrm{Pre}}]$ as the "placebo effect." These interpretations erroneously assume that had an individual not received the treatment or placebo, the measured outcome would not have changed over the duration of the study. Without a control group, one cannot separate treatment effects from natural variation, regression to the mean, or other temporal changes of the dependent variable. Even with gold-standard measurement tools, at least some random within-subject variation is inevitable over time. This variation can create the appearance of response heterogeneity when none exists or mask true treatment response differences. As Atkinson and Batterham (1) described, this often represents "random error masquerading as response differences." Although individual effects remain unobservable, an unbiased estimate of the average causal treatment effect can be obtained by comparing observed outcomes between groups (5).

Quantifying TRH involves estimating the standard deviation of individual treatment effects, denoted σ_D for a population and SD_D in a sample. However, the standard deviation of individual treatment effects, SD_D, cannot be directly computed from observed data in parallel-group trials because it requires knowing each participant's response to both conditions. The observed variation in outcomes or changes within treatment groups includes both true response heterogeneity and other sources of variation, particularly within-subject random variability between measurement timepoints. Using baseline or preintervention scores to calculate change scores does not eliminate this issue.

BACKGROUND ON THE STANDARD DEVIATION OF INDIVIDUAL RESPONSES

Given these points, we would like to address a recently popularized method for assessing TRH, referred to as the $\mathrm{SD}_{\mathrm{IR}}$ (1, 2, 6–8). The proponents of this specific method have claimed that *1*) it can be used to detect and estimate the degree of TRH, *2*) the estimated TRH represents TRH caused by the treatment, and *3*) the estimated TRH can be used to estimate the proportion of "responders." All three claims have important assumptions that, when violated, rise to the level of errors. Herein, we aim to explain these assumptions and how they can be addressed. However, we would like to note that this $\mathrm{SD}_{\mathrm{IR}}$ approach is inferentially superior to many other methods for investigating TRH. Simple responder counts by dichotomizing research participants as "responders" or "nonresponders" are presumptive, and the superiority of $\mathrm{SD}_{\mathrm{IR}}$ in this regard is clearly discussed by Atkinson et al. (7).

The $\mathrm{SD}_{\mathrm{IR}}$ method has been proposed for studies involving continuous outcomes that are measured both before and after receiving some treatment (pre-post measurements in parallel groups). The change scores for an outcome for every individual in the study can then be calculated. The logic proposed by proponents of the $\mathrm{SD}_{\mathrm{IR}}$ approach is the following: The standard deviation of the change scores only includes the within-subject variation (σ_{w}^2) and measurement error (σ_{ε}^2). Within-subject variation encompasses multiple sources, including biological fluctuations over time, behavioral changes unrelated to treatment, environmental factors, and the natural stability or instability of the outcome being measured. For example, in a study of blood pressure medications, σ_{w}^2 might include daily fluctuations in blood pressure due to stress, diet, sleep patterns, and seasonal effects, whereas in a fitness study measuring maximal oxygen consumption, it could include variations due to motivation during testing or circadian rhythms due to timing of the tests. Measurement error reflects the technical precision of the measurement tools. If treatment effects vary across individuals ($\sigma_{\mathrm{IR}}^2 > 0$), the treatment group will exhibit greater change score variability than the control group (i.e., $\sigma_{\mathrm{w}}^2 + \sigma_{\varepsilon}^2 + \sigma_{\mathrm{IR}}^2 > \sigma_{\mathrm{w}}^2 + \sigma_{\varepsilon}^2$). In simpler terms, the logic of $\mathrm{SD}_{\mathrm{IR}}$ is that if the treatment impacts people differently, the variability of change scores (improvements or declines) will be greater in the treatment group than in the control group. This assumption is based on the premise that both groups are equally

exposed to nontreatment-related factors affecting observed change scores (e.g., random within-person variation, measurement error, true temporal changes unrelated to treatment, etc.), and only the treatment group receives an active intervention. Thus, an observed difference in variances between groups is attributed to individual variation in the treatment response. For this logic to hold, the control group must function as a negative control, meaning it should exhibit no true interindividual response variance, and all observed variance can be attributed to within-subject variation and measurement error. This logic makes a critical mathematical assumption: individual treatment effects D are independent of the effects of the control (see Supplemental Material).

The SD_{IR} is calculated as:

$$SD_{IR} = \text{sgn}(s_T^2 - s_C^2)\sqrt{|s_T^2 - s_C^2|},$$

which relies on the estimated standard deviation of change scores in the treatment (s_T) and control groups (s_C). Confidence intervals can be formed on the differences in variances using a normal approximation, and the square root of those bounds can be taken to obtain the desired confidence limits (2, 9). The sgn function included in our equation for calculating SD_{IR} serves to always make SD_{IR} positive for situations in which s_T is less than s_C, when the difference would be negative. Barring the hassle of potential sign errors if the sgn function is not used, there is nothing statistically erroneous about calculating SD_{IR} itself because it is just an estimate of the square root of the difference in variances.

Some have advocated using SD_{IR} to estimate the proportion of responders (4, 8). They use the standard normal cumulative distribution function ($\Phi(\cdot)$) to estimate the proportion of individuals whose treatment effect would exceed a specified threshold (typically the minimally clinically important difference or the minimal detectable difference):

$$\text{Proportion of responders} = \Phi\left(\frac{\overline{D} - t}{SD_{IR}}\right),$$

where $\overline{D}$ is the mean difference between the treatment and control groups, and t is the threshold indicating the smallest difference of interest.

THE STATISTICAL ISSUES

Issue No. 1: The Fundamental Problem of Causal Inference

As some have illustrated (8), SD_{IR} seems akin to a random slope component from a mixed-effects model. As stated by Atkinson et al. (7), the SD_{IR} "is considered a parameter for the distribution of true responses in the population of interest alongside the mean treatment effect." This is a major assumption that might often be incorrect (see Supplemental Material for a greater discussion of this assumption). Although it is tempting to interpret this as being a way to estimate the variance of D, it relies on a critical assumption about something we cannot observe: the correlation between an individual's outcome with the treatment and their outcome with the control [i.e., the correlation between $Y(T)$ and $Y(C)$ for the same person]. To illustrate, imagine we could magically give the same person both treatment and control simultaneously. Would someone who has a favorable outcome with the treatment have a similarly favorable outcome with the control? This correlation is mathematically well-defined within the potential outcomes framework and directly affects variance calculations, but it remains fundamentally unobservable in parallel-group designs. This is often referred to as the "fundamental problem of causal inference" (10) because we can only observe a single potential outcome for each participant within a parallel-group experimental design. The SD_{IR} calculation implicitly assumes this correlation takes a specific value, but alternative correlation values could yield substantially different estimates of TRH.

When TRH is present, the genuine impact of treatment(s) varies among different individuals within a population. To clarify these points, we shall consider two potential outcomes, $Y(C)$ and $Y(T)$, from a two-arm parallel-group trial mentioned earlier. The variable $Y(T)_i$ is the value of the outcome on participant i when exposed to the treatment, and $Y(C)_i$ is the value if exposed to the control condition. The two values are imagined to be measured at the same moment in time. In a parallel-group trial, only the outcome corresponding to the group to which the participant was assigned is observable. The concept of two potential outcomes aids in conceptualizing a true treatment effect. Assuming other assumptions are met, the true treatment effect has a mean and variance defined by Gadbury and Iyer (11) as the following:

$$\overline{D} = \overline{Y(T)} - \overline{Y(C)}.$$

$$\sigma_D^2 = \sigma_T^2 + \sigma_C^2 - 2\sigma_T\sigma_C\rho_{TC}.$$

The marginal distributions of $Y(T)$ and $Y(C)$ are readily available from any parallel-group trial, but the correlation between potential outcomes, ρ_{TC}, is not (empirically) identifiable. For TRH to be absent ($\sigma_D = 0$), both the variance of $Y(C)$ and $Y(T)$ would have to be equal, and the correlation (ρ_{TC}) equal to 1. The central challenge in estimating TRH is the inability to estimate this correlation from real, observed data. For this reason, the presence of a difference in the variances between treatment conditions (at the population parameter level, not at the sample statistic level) is a sufficient but not necessary condition to indicate that TRH is present.

Because ρ_{TC} must lie within the interval $[-1, 1]$, upper and lower bound estimates of σ_D can be calculated. The lower bound, which occurs when $\rho_{TC} = 1$, for the sample variance is $\min(SD_D) = |s_T - s_C|$. The upper bound, which occurs when $\rho_{TC} = -1$, is $\max(SD_D) = s_T + s_C$. The standard deviation representing TRH could be greater than $\min(SD_D)$ but cannot exceed $\max(SD_D)$ (11). Furthermore, it could be reasonable to speculate that ρ_{TC} is closer to 1 than -1, but this is an untestable assumption.

Given the mathematical assumptions laid out by Gadbury et al. (12), it becomes clear that the SD_{IR} should not uncritically be assumed to be a "parameter [or unbiased estimate thereof] for the distribution of true responses." The use of SD_{IR} involves making strong assumptions that cannot be empirically tested because we cannot identify a correlation between potential outcomes (ρ_{TC}). Indeed, these assumptions may be untenable in many exercise science studies (see

Supplemental Material). Instead, one can loosen the strict assumptions made by SD_{IR} by presenting bounds of the SD_D. The strict assumptions made by SD_{IR} propagate into the calculation of the proportion of responders. If SD_{IR}'s strict assumptions are not met, the true denominator, the SD_D, could be considerably larger or smaller. The proportion of those who benefit ($P_{benefit} = Pr[D > 0]$), are harmed ($P_{harm} = Pr[D < 0]$), or even to respond beyond a predetermined threshold ($P_{responder} = Pr[D > t]$) can be estimated from the bounds on SD_D (12). Using these bounds and the estimate of $\overline{D}$, we can then use the standard normal cumulative distribution function to estimate these probabilities.

Issue No. 2: Sources of Individual Response Variance

Proponents of SD_{IR} have articulated that TRH is only present when the variance is greater in a treatment arm compared with the control arm. It was stated by Atkinson et al. (6) that, "In a parallel group study, true individual differences in exercise response are present only if the SD of change is substantially larger in the exercise group than the control group." Yet, it has been discussed elsewhere that TRH can exist even when the variance is not greater in the treatment group (9, 13). In fact, treatments could "homogenize" outcomes relative to the control group, which arguably still represents TRH (i.e., the degree of treatment effect still varies between individuals).

By extension, this implies that the intervention must cause a difference in population variances. This might be a plausible (but not necessary or verifiable) circumstance for very short placebo-controlled trials wherein an experimental drug might be the only major source of additional variance. However, such a broad assumption would certainly be too bold for comparative effectiveness studies that involve two competing treatments (e.g., new treatment vs. standard-of-care), wherein individuals may have varying responses to both treatments. A difference in variance does imply some degree of heterogeneity (though importantly, the converse and inverse are not true), and this conclusion should not be restricted to when the experimental group has a higher variance than the control group (14). Comparisons of variances between the treatment groups should then be two-sided and consider that there could be greater variance in either the control or treatment arm of a parallel-group trial.

Beyond this conceptual limitation, the SD_{IR} framework is also restricted in its applicability, as it can only be used in parallel-group trials where pre-post measurements are available to calculate change scores. In contrast, the SD_D and other heteroskedasticity tests provide a more versatile framework that can be applied to any parallel-group design, including those where only post-intervention measurements are available. This broader applicability is especially valuable in settings where baseline measurements are impractical, impossible, or simply were not collected in an existing dataset, allowing researchers to investigate TRH across a wider range of study designs (9).

Issue No. 3: Statistical Efficiency and Testing

Though the SD_{IR} calculation provides useful insights, it lacks statistical efficiency when comparing group variances, typically requiring substantial sample sizes for precise interval estimates. Statistical efficiency here refers to how effectively a method extracts information with minimal variance—essentially the precision of parameter estimation compared with alternative approaches. Furthermore, although variance comparisons across groups can help identify TRH, this methodology, whether using SD_{IR} or other techniques, faces substantial limitations.

Compared with other heteroskedasticity tests (e.g., variance ratio test), the SD_{IR} approach has less statistical power, with this effect being most pronounced at very small sample sizes, where even the nominal α level is not maintained (see Supplemental Material). Given that many studies in physiology research have small sample sizes, researchers may overlook possible TRH when further investigation is warranted (i.e., high type II error rates). In particular, differences in variances between groups may be hard to detect based on "statistical significance." For example, to achieve 80% statistical power when testing for heteroskedasticity between two groups, a study would need 68 participants per group to detect a twofold variance increase, and nearly 3,500 participants per group to detect a modest 10% difference in variances (15). Most physiological studies (16) and even meta-analyses lack sufficient statistical power to detect such differences in variances between treatment groups. Physiologists should exercise substantial caution when determining TRH presence based on heteroskedasticity tests with limited sample sizes (e.g., fewer than 60 participants per group, and certainly, they should not conclude the absence of TRH based on such tests—see Supplemental Material for more details).

Comparing variances to test for TRH dates back to a 1938 letter from Ronald A. Fisher to Henry Daniels (17), and various tests for detecting heteroskedasticity have been developed since (9). Although substantial variance differences between groups should warrant additional investigation into why individuals might respond differently to treatments, it is important to note that TRH does not always result in increased variance in the treatment compared with the control arm (9, 13). Additional caution is also warranted because heteroskedasticity may reflect measurement scale issues. For example, in some cases, a transformation (like the log-transformation) can eliminate heteroskedasticity (18), but the interpretation of TRH on the clinically meaningful scale should not be ignored [see Poulson et al. (19) for a greater discussion].

Even when group variances are equal, σ_D could be meaningfully large, indicating true individual differences in treatment response. Consider a scenario where some individuals experience increased benefit, whereas others experience decreased benefit from treatment. Such differential responses could produce similar overall variance between treatment and control groups while still generating substantial σ_D. Tests of heteroskedasticity could incorrectly be interpreted as suggesting no TRH exists in this case, because they only capture differences in group variances rather than the true variance of individual treatment effects. The variance components from a parallel-group trial cannot be used to definitively rule out the potential of TRH.

Therefore, although finding different variances between groups is sufficient to demonstrate heterogeneous treatment effects, it is not a necessary condition: TRH can exist even when variances are similar between groups. These tests may be useful when the intention is to establish sufficient conditions for claiming TRH, but the failure of these tests to reach statistical significance cannot be used to eliminate the possibility of TRH. Researchers should be cautious about relying solely on variance comparisons when making claims regarding TRH, especially in scenarios where sample sizes are small and tests of heteroskedasticity will be underpowered. This issue is obviously not limited to the SD_{IR}, but it is more pronounced compared with other, more efficient tests. Compared with other issues, the problem of efficiency for heteroskedasticity tests is a relatively minor issue that primarily affects those making dichotomous decisions based on *P* values or confidence intervals from heteroskedasticity tests. If researchers avoid making dichotomous decisions regarding the presence or absence of TRH, the seriousness of this issue is diminished.

Summary of the Issues

Although the SD_{IR} approach represents an advancement over simple so-called responder counts, it has several important limitations that researchers should consider that are highlighted in Table 1. First, the fundamental problem of causal inference means we cannot identify the correlation between potential outcomes, making it impossible to directly estimate the true standard deviation of individual treatment effects (σ_D). The SD_{IR} makes a strong assumption concerning the correlation between treatment and control outcomes; when this assumption is not thoughtfully justified, bounds would be more appropriate, even as a sensitivity analysis. Second, the assumption that TRH only exists when treatment group variance exceeds control group variance is overly restrictive. Heterogeneous treatment effects can exist even when variances are equal between groups or when the control group has greater variance. Third, the SD_{IR} lacks statistical efficiency compared with alternative methods like the variance ratio test, potentially leading to more underpowered analyses of TRH, especially in studies with modest sample sizes. Also, researchers should also be cautious in their interpretation of heteroskedasticity tests and how they relate to the presence or absence of TRH.

It is important to recognize that traditional parallel-group experimental designs, while efficacious for estimating average treatment effects, are not designed to investigate TRH. These designs were primarily developed to answer population-level questions about mean effects, not individual-level questions about response variability. Consequently, researchers should be cautious about making definitive claims regarding the presence or absence of TRH based solely on parallel-group trials and a heteroskedasticity test. The absence of statistically significant heteroskedasticity test as evidence for TRH should not be interpreted as definitive evidence that no meaningful TRH exists (only that there is a lack of evidence of TRH). Alternative designs that expose individuals to both treatment and control conditions, such as replicate crossover designs and variations on the Balaam design, may be better suited for investigating true individual differences in treatment response (20, 21).

AN EXAMPLE OF AN APPROPRIATE TRH ANALYSIS

Now, let us inspect a previously published study by Plotkin et al. (22) as an example of how to appropriately discuss and display TRH from a parallel-group trial. This study aimed to assess how muscle size and strength adaptations differ between two exercise prescription patterns, which aimed to progressively increase load or repetitions (LOAD and REPS, respectively). The SD_{IR} approach becomes particularly problematic when applied to comparative effectiveness studies such as these. In such designs, the fundamental assumption that the differential effects of

Table 1. *Summary of issues with the standard deviation of individual response*

Claim	Problem	Solution
"[SD_{IR}] is considered a parameter for the distribution of true responses in the population of interest alongside the mean treatment effect" (7)	• Makes a strong, rarely justified assumption concerning the correlation between potential outcomes. • We cannot (empirically) identify this correlation, making the assumption untestable.	• Present bounds for the SD_D (12) • These bounds could also be presented alongside response distributions and bounds for the probability of harm or benefit
"In a parallel group study, true individual differences in exercise response are present only if the SD of change is substantially larger in the exercise group than the control group" (6)	• TRH can exist when the control group shows greater variance or even when the group variances are equal.	• Ensure any comparisons of the variances are two-sided • Do not necessarily disregard the potential of TRH if the control group has greater or equal variance
SD_{IR} can be used to detect TRH by comparing variances between treatment groups	• SD_{IR} is an inefficient method. • Differences in variances are only a sufficient and not necessary condition to declare that there is nonzero TRH.	• Use other tests or estimates of heteroskedasticity; see Mills et al. (9) for possible suggestions • Do not conclude or claim that there is no presence of TRH based solely on a nonsignificant test of heteroskedasticity

SD_D, standard deviation of treatment effects; SD_{IR}, standard deviation of individual responses; TRH, treatment response heterogeneity.

the interventions can be represented by an orthogonal variance component is tenuous. Moreover, in studies comparing two exercise modalities (i.e., LOAD vs. REPS), designating one intervention as the "control" is arbitrary, as the SD_{IR} assumptions seem most valid in cases when there is a negative control (e.g., a placebo condition). Notably, proponents of the SD_{IR} approach have primarily focused their methodological discussions on classic intervention-versus-(negative) control designs, leaving a significant gap in guidance for comparative effectiveness research. Here, we outline an analysis that could be applied to studies parallel-group trials with or without a negative "control" group.

An analysis of covariance can be used to estimate the average treatment effect ($\overline{D}$). In the case of Plotkin et al. (22), we estimate the average treatment effect while adjusting for the preintervention one-repetition maximum back squat and the biological sex of the participant. This analysis indicated that the average treatment effect favored LOAD by 2.03 kg over REPS, 95% confidence interval [−3.99, 8.05].

The change score standard deviation estimates are 7.65 and 12.24 for the REPS and LOAD groups, respectively. Heteroskedasticity tests yield *P* values ranging from 0.0506 to 0.0775 (see Supplemental Material). Although none of these tests meet the traditional threshold for "statistical significance", notably, the standard deviation in the LOAD group is 1.6 times greater than that of REPS, which should (but alone is not needed to) prompt greater scrutiny of TRH. Because the study did not have a negative control, the labeling of which group is a control is arbitrary—the direction of heteroskedasticity (i.e., which treatment group has greater variance) should always be ignored in such trials.

From the observed change scores, the bounds of SD_D point estimates range from 4.6 to 19.9 kg, with considerably wide confidence intervals for these point estimates (Fig. 1*A*). Based on these point estimates and normality assumptions,

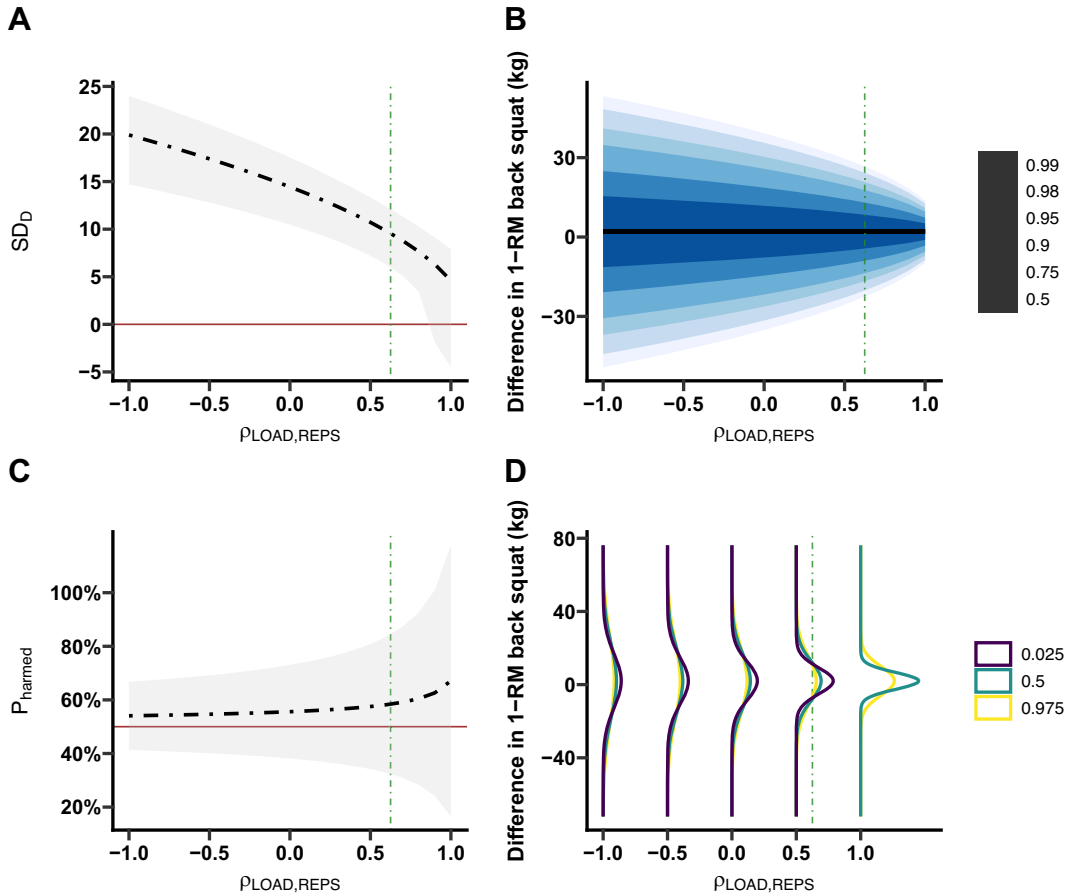

Figure 1. Relationship between the correlation of potential outcomes ($\rho_{LOAD,REPS}$) and estimated treatment response heterogeneity (TRH). *A*: TRH can be easily understood as the standard deviation of the treatment effect (SD_D), which varies as a function of the assumed correlation of potential outcomes. In the case of Plotkin et al. (22), the assumed correlation has drastic effects on the expected TRH, and importantly, there is much uncertainty in these estimates (black dotted line = point estimate; gray ribbon = 95% CI; red line = no effect). *B*: using the point estimates of SD_D, the expected distributions of treatment responses can be depicted (color = percentile and 1-percentile of the treatment response). This plot makes it apparent that, under certain assumptions, TRH dwarfs the average treatment effect. *C*: based on the treatment response distributions in *B*, the probability of harm (P_{harmed}) can be calculated (12). Note that there is considerable uncertainty associated with these point estimates, which incorporates the uncertainty of the intervention effect and the SD_D. *D*: since there is appreciable uncertainty in the estimation of SD_D, this can be incorporated into depictions of the response distributions. Whereas *B* depicts response distributions that use the point estimates of SD_D, *D* shows distributions for the point estimate (green) and 95% CI of SD_D (dark blue to yellow) for specific values of $\rho_{LOAD,REPS}$. In every panel, the dot-dash style green vertical line represents the value of $\rho_{LOAD,REPS}$ assumed by the SD_{IR}. 1-RM, one-repetition maximum; CI, confidence interval; SD_{IR}, standard deviation of individual responses.

we can graphically depict the expected treatment response distribution as a function of the assumed correlation ($\rho_{LOAD,REPS}$). From these individual treatment effect distributions, the probability of harm from the REPS treatment can be estimated. The bounds of P_{harmed} ranged from 54% to 67% with, again, considerable variability in these estimates due to the sample size (Fig. 1*C*). Using this information, we could then add that given the average treatment effect, and our estimates of SD_D, a sizeable number of individuals could have a net negative effect of using the REPS training compared with LOAD. Finally, because there is appreciable uncertainty in SD_D, we can depict how this uncertainty propagates into the expected distributions of treatment effects at certain values of $\rho_{LOAD,REPS}$ (Fig. 1*D*). Illustrating treatment response distributions provides readers with a clear and transparent picture of the implications of TRH across a range of assumptions that one can make, whereas SD_{IR} provides an estimate of TRH with a strict assumption concerning the value of $\rho_{LOAD,REPS}$ (vertical dot-dash green lines).

The authors could then conclude that given the small effect size, especially relative to our different sources of variance, neither treatment is likely to confer much benefit or harm compared with the other. However, given the potential bounds of SD_D, the authors may consider the discussion of TRH worthy of further inspection. Potential interactions could be explored, and covariates could be used to tighten the bounds on SD_D (12, 21). The authors could also provide some speculations concerning unmeasured factors influencing TRH and propose study designs that could address these factors in the future (21). Finally, researchers should avoid using TRH as a default explanation when results are ambiguous or fail to show "statistically significant" effects. Attributing such results to TRH without theoretical justification is not a rigorous interpretation of parallel-group trials.

CONCLUSIONS

The investigation of TRH represents an important frontier in understanding how interventions affect different individuals, particularly as the field moves toward personalized medicine. Although the SD_{IR} method improved upon other approaches, it makes strong assumptions that restrict its utility for estimating true TRH in many parallel-group trials, especially comparative effectiveness trials, which are ubiquitous in exercise physiology. Most importantly, the SD_{IR} approach cannot resolve the fundamental problem of causal inference, potentially making erroneous assumptions when attempting to reflect true individual treatment effect variance. This approach's strict requirements might only be viable under specific circumstances (detailed in Supplemental Material)—researchers continuing to use this method should therefore justify how their trial satisfies these necessary conditions.

As Rothman et al. (23) astutely observed, disagreements over TRH often stem from failure to separate distinct contexts in which heterogeneity is evaluated: statistical, biological, public health, and individual decision-making. Each context has different implications for how treatment heterogeneity should be assessed and discussed. The SD_{IR} approach primarily operates in the statistical context, and the other three contexts that are crucial for meaningful translation of research findings.

Rather than relying solely on SD_{IR}, we recommend that researchers think carefully about their goals and use analyses that address those goals:

1) If researchers want to establish the presence of TRH, then evidence of the sufficient condition of heteroskedasticity should be enough. Although such tests are sufficient to evidence the presence of TRH, they cannot rule it out, even if they have a very precise estimate that indicates homoskedasticity.
2) If researchers are interested in estimating the degree of TRH, they are left with SD_{IR} and SD_D. SD_{IR} makes strong assumptions about ρ_{TC}, which, in many cases, are tenuous. Thus, the safest option is to present SD_D with the relevant graphical depictions, and perhaps the authors can comment on which ρ_{TC} may be more or less reasonable depending on their experiment and theory.
3) Estimate bounds for the probability of harm or benefit or from treatments, which provides more actionable information for clinical decision-making and better aligns with what Rothman et al. (23) identified as essential for both public health and individual contexts.
4) When a salient degree of heterogeneity is detected, carefully investigate potential mechanisms and moderators rather than simply attributing results to undefined "individual differences."

Importantly, researchers should avoid using TRH as a default explanation for "nonsignificant" differences without a theoretical justification. Asserting that TRH is an explanation, despite negligible average effects, necessarily implies "harm" to certain subgroups—an assumption that warrants scrutiny. The field of exercise physiology would benefit from moving beyond simply detecting heterogeneity and toward understanding its underlying causes and implications for treatment decisions (21). Of particular concern is the widespread practice of classifying individuals as "responders" and "nonresponders" based on observed changes, which represents what Senn (24) termed "dichotomania"—the problematic transmutation of continuous outcome measures into binary classifications. This practice can be misleading due to the fundamental problem of causal inference that we described earlier and substantial losses in statistical efficiency and interpretability. Rather than speciously pursuing responder identification in small studies, researchers should prioritize collecting large, diverse samples to detect meaningful moderators of treatment effects. In addition, modern advances in experimental design offer many improvements over the parallel-group design for the identification of TRH (25), exemplified by innovative approaches such as replicated within-participant unilateral trials that can more effectively separate true individual differences from measurement error and biological variability (26). This approach shift is imperative to avoid perpetuating analyses that may reflect statistical artifacts rather than true individual variation, and in turn, it will facilitate genuine personalized medicine and exercise prescriptions.

These recommendations aim to advance more rigorous investigation of TRH while acknowledging the inherent

limitations in parallel-group designs. By adopting more nuanced analytical approaches that explicitly address statistical, biological, public health, and individual decision-making contexts, and by maintaining high standards for causal inference, researchers can better contribute to the development of truly personalized treatment approaches.

SUPPLEMENTAL MATERIAL

Supplemental Material: https://doi.org/10.5281/zenodo.15492100.

ACKNOWLEDGMENTS

We thank Dan Beavers and Stephen Senn for feedback in discussions about treatment response heterogeneity and feedback on early drafts of this manuscript.

DISCLOSURES

Collectively, the authors and their institutions have received payments for consultations, grants, contracts, in-kind donations, and contributions from multiple for-profit and not-for-profit entities interested in statistical design and analysis of experiments but not directly related to the methodological questions addressed in the present paper. A.W.B. is currently supported by associated NIH Grants: R25HL124208; R25DK099080; and P20GM109096. The views expressed in the manuscript do not necessarily reflect the views of the funders or any other organization.

AUTHOR CONTRIBUTIONS

A.R.C. prepared figures; A.R.C., D.B.A., A.W.B., G.L.G., T.R.M., R.D.S., and A.D.V. drafted manuscript; A.R.C., D.B.A., A.W.B., G.L.G., T.R.M., R.D.S., and A.D.V. edited and revised manuscript; A.R.C., D.B.A., A.W.B., G.L.G., T.R.M., R.D.S., and A.D.V. approved final version of manuscript.

REFERENCES

1. **Atkinson G**, **Batterham AM.** True and false interindividual differences in the physiological response to an intervention. *Exp Physiol* 100: 577–588, 2015. doi:10.1113/EP085070.
2. **Hopkins WG.** Individual responses made easy. *J Appl Physiol (1985)* 118: 1444–1446, 2015. doi:10.1152/japplphysiol.00098.2015.
3. **Bonafiglia JT**, **Preobrazenski N**, **Gurd BJ.** A systematic review examining the approaches used to estimate interindividual differences in trainability and classify individual responses to exercise training. *Front Physiol* 12: 665044, 2021. doi:10.3389/fphys.2021.665044.
4. **Swinton PA**, **Hemingway BS**, **Saunders B**, **Gualano B**, **Dolan E.** A statistical framework to interpret individual response to intervention: paving the way for personalized nutrition and exercise prescription. *Front Nutr* 5: 41, 2018. doi:10.3389/fnut.2018.00041.
5. **Laird N.** Further comparative analyses of pretest-posttest research designs. *Am Stat* 37: 329–330, 1983. doi:10.1080/00031305.1983.10483133.
6. **Atkinson G**, **Williamson P**, **Batterham AM.** Exercise training response heterogeneity: statistical insights. *Diabetologia* 61: 496–497, 2018. doi:10.1007/s00125-017-4501-2.
7. **Atkinson G**, **Williamson P**, **Batterham AM.** Issues in the determination of 'responders' and 'non-responders' in physiological research. *Exp Physiol* 104: 1215–1225, 2019. doi:10.1113/EP087712.
8. **Swinton P.** Assessing individual response to training in sport and exercise (Preprint). *SportRxiv*, 2023. doi:10.51224/srxiv.288.
9. **Mills HL**, **Higgins JPT**, **Morris RW**, **Kessler D**, **Heron J**, **Wiles N**, **Davey Smith G**, **Tilling K.** Detecting heterogeneity of intervention effects using analysis and meta-analysis of differences in variance between trial arms. *Epidemiology (Fairfax)* 32: 846–854, 2021. doi:10.1097/EDE.0000000000001401.
10. **Holland PW.** Statistics and causal inference. *J Am Stat Assoc* 81: 945–960, 1986. doi:10.1080/01621459.1986.10478354.
11. **Gadbury GL**, **Iyer HK.** Unit-treatment interaction and its practical consequences. *Biometrics* 56: 882–885, 2000. doi:10.1111/j.0006-341x.2000.00882.x.
12. **Gadbury GL**, **Iyer HK**, **Allison DB.** Evaluating subject-treatment interaction when comparing two treatments. *J Biopharm Stat* 11: 313–333, 2001. doi:10.1081/bip-120008851.
13. **Cortés J**, **González JA**, **Medina MN**, **Vogler M**, **Vilaró M**, **Elmore M**, **Senn SJ**, **Campbell M**, **Cobo E.** Does evidence support the high expectations placed in precision medicine? A bibliographic review. *F1000Res* 7: 30, 2019. doi:10.12688/f1000research.13490.4.
14. **Senn S.** Mastering variation: variance components and personalised medicine. *Stat Med* 35: 966–977, 2016. doi:10.1002/sim.6739.
15. **NCSS Statistical Software.** PASS 2024 Power Analysis and Sample Size (PASS) (Online). The United States Department of Veterans Affairs, 2024.
16. **Renwick JRM**, **Preobrazenski N**, **Wu Z**, **Khansari A**, **LeBouedec MA**, **Nuttall JMG**, **Bancroft KR**, **Simpson-Stairs N**, **Swinton PA**, **Gurd BJ.** Standard deviation of individual response for VO2max following exercise interventions: a systematic review and meta-analysis. *Sports Med* 54: 3069–3080, 2024. doi:10.1007/s40279-024-02089-y.
17. **Fisher RA.** *Statistical Inference and Analysis: Selected Correspondence of RA Fisher*. Oxford University Press, 1990.
18. **Senn S.** Controversies concerning randomization and additivity in clinical trials. *Stat Med* 23: 3729–3753, 2004. doi:10.1002/sim.2074.
19. **Poulson RS**, **Gadbury GL**, **Allison DB.** Treatment heterogeneity and individual qualitative interaction. *Am Stat* 66: 16–24, 2012. doi:10.1080/00031305.2012.671724.
20. **Senn S**, **Rolfe K**, **Julious SA.** Investigating variability in patient response to treatment–a case study from a replicate cross-over study. *Stat Methods Med Res* 20: 657–666, 2011. doi:10.1177/0962280210379174.
21. **Zoh RS**, **Esteves BH**, **Yu X**, **Fairchild AJ**, **Vazquez AI**, **Chapple AG**, **Brown AW**, **George B**, **Gordon D**, **Landsittel D**, **Gadbury GL**, **Pavela G**, **de Los Campos G**, **Mestre LM**, **Allison DB.** Design, analysis, and interpretation of treatment response heterogeneity in personalized nutrition and obesity treatment research. *Obes Rev* 24: e13635, 2023. doi:10.1111/obr.13635.
22. **Plotkin D**, **Coleman M**, **Van Every D**, **Maldonado J**, **Oberlin D**, **Israetel M**, **Feather J**, **Alto A**, **Vigotsky AD**, **Schoenfeld BJ.** Progressive overload without progressing load? The effects of load or repetition progression on muscular adaptations. *PeerJ* 10: e14142, 2022. doi:10.7717/peerj.14142.
23. **Rothman KJ**, **Greenland S**, **Walker AM.** Concepts of interaction. *Am J Epidemiol* 112: 467–470, 1980. doi:10.1093/oxfordjournals.aje.a113015.
24. **Senn S.** Dichotomania: an obsessive compulsive disorder that is badly affecting the quality of analysis of pharmaceutical trials. *Proceedings of the International Statistical Institute, 55th Session*. Sydney, 2005.
25. **Robinson ZP**, **Helms ER**, **Trexler ET**, **Steele J**, **Hall ME**, **Huang C-J**, **Zourdos MC.** N of 1: optimizing methodology for the detection of individual response variation in resistance training. *Sports Med* 54: 1979–1990, 2024. doi:10.1007/s40279-024-02050-z.
26. **Robinson ZP**, **Steele J**, **Helms ER**, **Trexler ET**, **Hall ME**, **Huang C-J**, **Pelland JC**, **Remmert JF**, **Hinson SR**, **Mikula SA**, **Hamaïde AA**, **Zourdos MC.** The effect of resistance training volume on individual-level skeletal muscle adaptations: a novel replicated within-participant unilateral trial (Preprint). *bioRxiv*, 2025. doi:10.1101/2025.07.24.666533.