Perspectives on Behavior Science (2022) 45:667–679
https://doi.org/10.1007/s40614-022-00346-x

SI: COMMENTARY ON SLOCUM ET AL, THREATS TO INTERNAL VALIDITY

Commentary on Slocum et al. (2022): Additional Considerations for Evaluating Experimental Control

Sean W. Smith[1] · **Faris R. Kronfli**[1] · **Timothy R. Vollmer**[1]

Accepted: 6 June 2022 / Published online: 21 July 2022

Abstract

In the target article, Slocum et al. (2022) suggested that nonconcurrent multiple baseline designs can provide internal validity comparable to concurrent multiple baseline designs. We provide further support for this assertion; however, we highlight additional considerations for determining the relative strength of each design. We advocate for a more nuanced approach to evaluating design strength and less reliance on strict adherence to a specific set of rules because the details of the design only matter insofar as they help researchers convince others that the results are valid and accurate. We provide further support for Slocum et al.'s argument by emphasizing the relatively low probability that within-tier comparisons would fail to identify confounds. We also extend this logic to suggest that staggering implementation of the independent variable across tiers may be an unnecessary design feature in certain cases. In addition, we provide an argument that nonconcurrent multiple baseline designs may provide verification within baseline logic contrary to arguments made by previous researchers. Despite our general support for Slocum et al.'s assertions and our advocacy for more nuanced approaches to determining the strength of experimental designs, we urge experimenters to consider the perspectives of researchers from other fields who may favor concurrent multiple-baseline designs and suggest that using concurrent multiple-baseline designs when feasible may foster dissemination of behavior analytic research.

Keywords Multiple-baseline design · Concurrent · Nonconcurrent

✉ Sean W. Smith
seansmith1@ufl.edu

[1] Department of Psychology, University of Florida, 945 Center Drive, Gainesville, FL 32611-2250, USA

Introduction

In the target article, Slocum et al. (2022) provided a compelling argument that nonconcurrent multiple-baseline designs (NonconMBLs) can demonstrate internal validity comparable to concurrent multiple-baseline designs (ConMBLs). Slocum et al. outlined various threats to internal validity and cogently explained how each design controls for potential confounds with either within- or across-tier analyses. The authors explained how NonconMBLs and ConMBLs only differ based on the extent to which they allow for across-tier analyses (i.e., ConMBLs allow across-tier analyses, NonconMBLs do not), but they argued that the putative increase in internal validity provided by across-tier analyses is based on somewhat paradoxical logic. They explained that the effects of an independent variable (IV) must not generalize across tiers in a multiple-baseline design to demonstrate experimental control—the dependent variable (DV) should change when and only when the IV is implemented. However, to identify a confound using across-tier analyses, the effects of an extraneous variable would have to generalize and affect the DV in a similar way across tiers. The authors argued that it is paradoxical to assume the effects of one class of variables (i.e., an extraneous variable) would generalize across tiers when another variable (i.e., the IV) clearly does not generalize across tiers. If the IV does not have an effect across tiers, it seems unlikely that other variables would have an effect across tiers, suggesting that identifying extraneous variables using across-tier comparisons is unlikely. Thus, the across-tier analyses that can be conducted within ConMBLs offer negligible increases in internal validity, suggesting that ConMBLs do not provide meaningful increases in internal validity compared to NonconMBLs even though the latter does not permit across-tier analyses. In general, we agree with this logic and the conclusion that NonconMBLs provide similar internal validity compared to ConMBLs.

We expand on Slocum et al.'s (2022) argument in several ways. First, we provide additional perspective to the discussion by highlighting that the strength of an experimental design is determined by a variety of variables—internal validity is only one factor to consider. NonconMBLs and ConMBLs differ along other dimensions of experimental control that researchers must consider when designing and evaluating experiments, which suggests that researchers should limit their reliance on strict adherence to specific sets of rules for determining the strength of an experimental design. Second, we supplement Slocum et al.'s argument against the "primary methodological criticism" of NonconMBLs (i.e., lack of across-tier comparisons) by describing the simple probabilities that within-tier comparisons would fail to identify confounds. Further, these simple probabilities are so small that we suggest it may be unnecessary to stagger implementation of the IV across tiers in certain experiments. Third, we highlight the potentially faulty logic supporting the "second methodological criticism" of NonconMBL (i.e., lack of "verification" in baseline logic) to further support Slocum et al.'s argument. Finally, despite our general support for Slocum et al.'s assertions, we consider alternative perspectives that may suggest favoring ConMBLs for reasons

other than the relative strength of each design, and we caution researchers against exclusively accepting a single point of view within this discussion.

Additional Perspectives on the Discussion of Internal Validity

Slocum et al. (2022) focused their discussion on the degree of internal validity offered by ConMBLs and NonconMBLs; however, researchers must evaluate experiments and their implications along other dimensions (e.g., external validity) as well. We highlight this to prevent readers from making a priori determinations regarding the relative strength of ConMBLs and NonconMBLs. Science is a social endeavor and the goal of any given experiment is to convince other scientists that the IV(s) caused an effect on the DV(s). Although the scientific community has developed certain guidelines that help researchers convince others of the effects demonstrated in their experiments, it is important to remember that these guidelines are not strict rules. There is nuance behind every guideline, and deviations from guidelines should be permitted as long as the experiment provides sufficient evidence to convince others of the effect.

No experimental design is perfect. Every experimental design has relative merits and limitations, and researchers must consider all of these when designing and evaluating experiments. In the context of discussing ConMBLs and NonconMBLs, some particularly relevant factors to consider are (1) the purpose of the experiment; (2) the likelihood of specific types of confounds based on the parameters of the experiment (e.g., DVs and IVs); (3) the method for addressing confounds (i.e., identifying vs. avoiding confounds); and (4) the patterns of data produced within the design. We highlight these considerations both to provide nuance to the discussion of ConMBLs and NonconMBLs and to reduce researchers' strict adherence to traditional rules governing implementation of research designs in general.

Purpose of the Experiment

There are often trade-offs to consider with experimental designs, and the relative value of each trade-off should be evaluated based on the purpose of the experiment. A good example of this consideration is elucidated in Ghaemmaghami et al.'s (2021) discussion of the efficacy and effectiveness of functional communication training. Designs that emphasize internal validity are warranted in studies evaluating the efficacy of an intervention if the effect of the IV on the DV does not have extensive empirical support. In contrast, research demonstrating effectiveness is evaluating the extent to which a well-documented effect remains intact as other variables change (e.g., settings, participants, treatment implementers), so tightly controlling many variables to enhance internal validity may limit an experiment's generality and undermine the overarching purpose of the study. In this way, external validity may take precedent over internal validity in an experiment evaluating effectiveness.

Considerations of internal and external validity may affect how one evaluates the relative strengths and weaknesses of ConMBLs and NonconMBLs. For example,

Slocum et al. (2022) discussed how across-tier comparisons enhance internal validity by selecting tiers that differ by only one factor (p. 14); however, this will likely decrease the external validity of the experiment because it limits the extent to which one can conclude that the IV will have similar effects when applied in increasingly dissimilar situations. Thus, when an experimenter wants to demonstrate efficacy, a ConMBL that can enhance across-tier comparisons may be advantageous. NonconMBLs may be more advantageous for an effectiveness study by showing that similar effects occur in tiers that are more isolated, which may permit more variables to differ across tiers, provide a stronger demonstration of generality, and give experimenters more flexibility in implementing their experiment with the resources they have available.

Likelihood of Specific Confounds

It is not possible for an experimental design to rule out all potential confounds. An experimental design simply needs to rule out confounds to such an extent that the researcher can convince others that the experimental effect was caused by the variables they manipulated. It should be noted that the likelihood that specific extraneous variables will arise differs across experiments, so design strength must be considered in terms of whether the design rules out the effects of the extraneous variables that are most likely to occur within the particular experiment. For example, in a study evaluating the efficacy of behavioral skills training (BST) for increasing the percentage of correct functional analysis (FA) components a therapist implements (e.g., Lloveras et al., 2022), it is unlikely that a single coincidental event could increase correct FA implementation across participants (with the exception of attending a separate FA implementation workshop, which could just as easily function as an exclusionary criterion for the study). In such a study, the across-tier comparison would not significantly enhance the internal validity of the study, and a NonconMBL would provide essentially the same degree of experimental control as a ConMBL. On the other hand, in a study evaluating the effects of differential reinforcement of alternative behavior for increasing appropriate behavior with a single participant across settings (with each setting represented as a separate tier), it is much more likely that a single coincidental effect (e.g., resolution of a sickness) could have the same effect across tiers in the same direction as the IV. Thus, a ConMBL would increase the internal validity of this study relative to a NonconMBL. Because the likelihood that a single confound can affect multiple tiers in an experiment depends on a multitude of factors, those factors must be considered when evaluating the overall strength of the design, and researchers should be appropriately cautious when making assertions about the strengths of designs a priori.

Identifying or Avoiding Confounds

ConMBLs and NonconMBLs may have different advantages based on *how* they establish experimental control. Slocum et al. (2022) described how ConMBLs permit across-tier comparisons, so the common conceptualization is that they will be

more likely to *identify* confounds caused by a single coincidental event (e.g., Carr, 2005; Gast et al., 2018; Harvey et al., 2004; Johnston et al., 2020). Slocum et al. also described how NonconMBLs may permit tiers that are more isolated, so a single coincidental event may be less likely to affect multiple tiers. In this way, NonconMBLs may be more likely to *avoid* having a single confound affect multiple tiers. Of course, it is good to identify confounds if they do occur, but it may be similarly beneficial to avoid such confounds from the outset.

For example, in experiments with nonhuman animals, the reinforcing efficacy of contingent food delivery will fluctuate depending on when the animal last ate. To *avoid* the influence of this confound, researchers often feed the animals such that they remain at less than 100% of their free-feeding body weight to ensure the animal remains hungry and motivated to engage in the behavior that produces access to food throughout the experiment (e.g., McDevitt et al., 2022). Although a researcher could allow this variable to influence responding and identify whether it could be a confound, it seems clear that avoiding this confound from the outset increases experimental control. Likewise, the increased isolation of tiers in NonconMBLs may allow researchers to avoid having a single coincidental variable affect multiple tiers, which further reduces the concern that NonconMBLs may be slightly less likely to identify across-tier effects.

Patterns of Behavior Change

The strength of a design is inherently related to the pattern of behavior change produced within the design. Although an experimental design may arrange conditions so clear patterns can emerge in the data, it does not matter what design or control procedures an experimenter implements if the experiment fails to produce those patterns. An experimental design will only convince other researchers of the experimental effect if the design actually produces patterns of behavior change that lead to confident conclusions regarding the relation between the IV and DV. Thus, a clear link between behavior change and the intervention contribute to the internal validity of the experiment as well. This is relevant to the discussion of ConMBLs and NonconMBLs because the patterns of data produced in the experiment influence conclusions about experimental control beyond whether the baselines are conducted concurrently.

In general, deviations from expected patterns of data (based on the nature of the relation between the DV and IV, which may have been established in previous research) may suggest the influence of an extraneous variable. For example, increasing or decreasing trends in behavior, abrupt changes without an intervention, or no change after an intervention, may suggest an extraneous influence. If data patterns deviate from what is expected, a design that rules out more extraneous variables could be more convincing than a design that rules out fewer variables. However, if (1) the data patterns are generally consistent with what is expected based on previous research; (2) implementation of the IV is reliably followed by a consistent change in the DV; (3) and this sort of change in the DV does not occur at other points in the experiment, a design that rules out more extraneous variables will not appreciably increase one's confidence relative to a design that may not rule out quite as many extraneous variables (especially if those extraneous variables are unlikely).

For example, if previous research suggests the IV should produce a large and immediate change in the DV, observing this pattern of data within a NonconMBL would be similarly convincing as observing this pattern within a ConMBL. This is because across-tier comparisons only provide small increases in experimental control, and these increases are negligible when the data show a clear relation. On the other hand, if the IV is expected to produce a large, immediate effect but the data patterns deviate from this expected pattern, then it is more likely that an extraneous variable could have influenced the DV. In this case (when the experimental effect is less clear based on the pattern of the data), the additional across-tier comparison provided by a ConMBL may help convince others of the experimental effect by ruling out slightly more confounds than a NonconMBL.

A Nuanced Perspective

These considerations combine to suggest that the relative strength of any given design cannot be assessed in the absence of other variables. Some of the variables that might be particularly relevant to the comparison of ConMBLs and NonconMBLs are (1) the purpose of the study; (2) the nature of the DVs and IVs; (3) whether it may be better to identify or avoid potential confounds; and (4) the patterns of the data produced within the experiment. We encourage consumers of research to avoid focusing on whether an experiment rigidly adheres to a specific set of rules and instead focus on the extent to which an experiment provides convincing evidence that the IV causes the change in the DV and rules out the possibility that other variables contributed to the change.

With this caveat, we supplement Slocum et al.'s (2022) assertion that NonconMBLs can often demonstrate the same experimental control as ConMBLs in the next sections by (1) highlighting the extremely low probability of failing to identify a confound using within-tier comparisons and (2) providing an argument that baseline logic can apply to NonconMBLs.

Necessity of Across-Tier Comparisons

A common criticism of NonconMBLs is that the lack of synchronized tiers precludes experimental control over potential confounds across time, which are often categorized as maturation, testing, and history/coincidental effects. Slocum et al. (2022) explained that within-tier comparisons control for maturation and testing effects, so there is no difference between ConMBLs and NonconMBLs when controlling for these confounds. Next, they explain that although within-tier comparisons do not completely control for coincidental effects, across-tier comparisons are unlikely to identify confounds caused by coincidental events beyond what would be evident with within-tier comparisons. We agree with their arguments, and in the section below we supplement their arguments by providing a description of the control provided by within-tier analyses based on simple probabilities. With this additional evidence, we extend Slocum et al.'s argument to suggest that the amount of control

provided by within-tier comparisons may obviate not only the need for concurrence, but also the need for staggering IV implementation across tiers.

Probability-Based Justification for Nonconcurrence

Christ (2007) demonstrated that, from a purely probability-based perspective, within-tier comparisons can provide sufficient experimental control for multiple baseline designs. Christ argued that it is improbable that the effect of an extraneous variable on a DV would coincide with the implementation of the IV across nonconcurrent tiers,[1] which makes it unlikely that the effect of an extraneous variable would be falsely attributed to the IV. To briefly summarize Christ's argument, the simple probability that an extraneous variable could cause a change in the DV at the same time the IV is implemented (and therefore be responsible for the change in the DV rather than the IV causing the change) is ultimately a function of (1) the number of occasions the experimenter collects data and (2) the number of times the IV is implemented. As the experimenter increases the number of occasions of data collection (i.e., the total number of data points in the experiment), the probability that the effect of an extraneous variable would only occur simultaneously with the implementation of the IV decreases. Although the probability that the effect of an extraneous variable coincides with an IV implementation increases as the experimenter increases the number of times they implement an IV, this is offset by the addition of more data points and replications of the experimental effect. Thus, the addition of data points within a single tier and the addition of tiers, each of which brings more data points and replications, systematically reduces the probability that the effect of an extraneous variable would coincide with and be misattributed to the IV implementation.

Christ (2007) operationalizes this logic in Table 2 of their article (which we have adapted and present as Table 1), which outlines the simple probabilities that an extraneous variable would coincide with the change in the DV as a function of different numbers of data points in each tier and the total number of tiers in a multiple baseline design. It should be noted that the probability that an extraneous variable would only cause changes that coincide with each implementation of the IV in an experiment with two tiers and six data points in each tier is .04, which seems like a sufficiently low probability to conclude that an extraneous variable did not cause each change in the DV. Yet, commonly accepted guidelines for single-subject research designs suggest more stringent parameters for the number of tiers and data points within each tier (e.g., Gast et al., 2018; Horner et al., 2005; Kazdin, 2021; Kratochwill et al., 2013; Slocum et al., 2022). In fact, the commonly accepted standards suggest a large enough number of data points and tiers to make the probability of extraneous variables causing every change in the IV negligible. It should be noted that this probability-based argument does not assume concurrence of tiers. This

[1] Hayes (1981) made a similar claim that a series of coincidences across nonconcurrent tiers is simply unlikely. We focus on Christ's (2007) argument because it provided exact probabilities and the variables controlling the probabilities.

Table 1 Probability that Effects of an Extraneous Variable will Coincide with One or More Instance of Independent Variable Implementations in Multiple Baseline Designs

Measurements		Tiers within Multiple Baseline Design					
Data Points per Tier	Intervals between Data Points	1	2	3	4	5	6
6	5	.2000	.0400	.0080	.0016	.0003	.0001
9	8	.1250	.0156	.0020	.0002	.0000	.0000
12	11	.0909	.0083	.0008	.0001	.0000	.0000
15	14	.0714	.0051	.0004	.0000	.0000	.0000
18	17	.0588	.0035	.0002	.0000	.0000	.0000
21	20	.0500	.0025	.0001	.0000	.0000	.0000
24	23	.0435	.0019	.0001	.0000	.0000	.0000

Based on Table 2 in Christ (2007).

suggests that, given a sufficient number of replications and data points, the simple probability of an extraneous variable causing the change in the DV that coincides with the IV implementation is so negligible that there may be no need for tiers to be implemented concurrently to be confident that the IV (and not extraneous variables) caused the change in the DV.

Although this argument does not assume concurrence of tiers, experimenters should implement the IV at different points in chronological time across each tier to reduce the probability that a single coincidental event could confound multiple tiers. Simultaneous implementation of the IV across tiers would increase the possibility that a single coincidental event would coincide with the IV implementation across tiers and confound the results of multiple tiers. Thus, implementing the IV at different points in time across tiers reduces the probability that the effects of an extraneous variable would be misattributed to the IV.

Argument Against the Need for Staggering Baselines

Extending this probability-based argument further, researchers could conceivably increase the number of tiers, the number of data points in each tier, or both to produce such low probabilities of extraneous variables causing each change in the DV that researchers may not need to stagger the number of *sessions* prior to IV implementation across tiers. The argument based on simple probabilities suggested by Christ (2007) does not rely on the assumption that there are a *different* number of sessions prior to IV implementation in each tier. Simply increasing the number of data points and tiers reduces the probability of extraneous variables causing a change in the DV that coincides with each IV implementation. When an experiment replicates the experimental effect across tiers numerous times and there is a relatively large number of data points in each tier, the likelihood that an extraneous variable would cause the change in the DV each time the IV is implemented becomes so low that it should convince others that the IV did in fact cause the change in the DV,

irrespective of the number of data points occurring before or after IV implementation across tiers.

Similar to the previous argument, the demonstration of the experimental effect will be most convincing when an experimenter implements the IV across tiers at different points in chronological time. It is important to note that chronological time and the number of sessions does not necessarily correspond. For example, in the first tier, 10 sessions could occur within 2 days, but in the second tier, 10 sessions could occur across 10 days. We argue that the experimenters can implement the IV after five sessions in both tiers, but the experimenter should implement the IV on different calendar days. Implementing the IV on different calendar days decreases the possibility that a single coincidental event would occur simultaneously with IV implementation across multiple tiers, so implementing the IV at different points in chronological time will provide the most convincing demonstration of control for the effects of coincidental events. However, experimenters need not vary the number of sessions before or after IV implementation because coincidental events correspond to chronological time, not to the number of sessions the experimenter conducts. Thus, to control for the effects of coincidental events, an experimenter only needs to vary IV implementation across chronological time, not the number of sessions.

We anticipate that this assertion will be unsettling to many researchers. A primary counterargument may be that, as pointed out by Slocum et al. (2022), the lag in IV implementation is the primary way for ruling out confounds related to maturation and testing. However, as we noted above, it is never possible to rule out all potential confounds, so an experimental design should only need to rule out confounds that are likely to affect a given experiment based on the nature of the IV, DV, or both, to convince someone of an experimental effect. Returning to our example of using BST to implement FA procedures, it is simply unlikely that a person will begin to implement FA procedures correctly due to maturation or testing effects (assuming the baseline condition does not include feedback or reinforcement). In such an experiment, it would be a waste of the experimenters' resources and the participants' time to stagger the implementation of the IV to rule out confounds that are highly unlikely. Thus, it seems reasonable to exclude this design feature given that it would not appreciably increase one's confidence that it was BST that caused the increase in correct FA implementation.

Lehardy et al. (2021) provided an excellent example of this logic. They systematically replicated the effect of a video-modelling intervention to teach 24 participants to create publication-quality graphs in Microsoft Excel. Although they did not stagger the implementation of their IV across participants, numerous other factors (e.g., a consistent and clear demonstration of an effect that was also consistent with previous research, unlikely contributions of maturation or testing effects, large number of data points, large number of replications, implementation of procedures like recording computer screens to rule out use of extra-experimental materials) made their demonstration convincing despite their deviation from traditional experimental strictures. We believe this experiment serves as a model for how staggering IV implementation may be an unnecessary design feature depending on a multitude of variables. It also emphasizes our more general argument that researchers should not be expected to follow rigid rules governing experimental designs, nor should

reviewers hold researchers to these arbitrary standards if the experimenters' deviations are justified. Instead, both researchers and reviewers should focus on whether the experiment provides a convincing demonstration of the effect of the IV on the DV based on the parameters of the specific experiment.

Verification Within NonconMBLs

Slocum et al. (2022) noted that a second major methodological criticism of NonconMBLs is levied by Cooper et al. (2020), who suggested that NonconMBLs do not fulfill the "verification" step in the traditional conceptualization of baseline logic because NonconMBLs lack real-time, across-tier comparisons. We agree with the alternative form of experimental logic that Slocum et al. suggested (i.e., prediction, contradiction, and replication); however, we also argue that verification can occur within NonconMBLs according to the traditional conceptualization of baseline logic.

With the traditional conceptualization of baseline logic, verification is established in a multiple baseline design when an IV causes a change to the DV in one tier, yet stable responding remains unchanged in a second tier. This demonstration verifies the prediction that the DV would not have changed if the IV had not been implemented.[2] Cooper et al. (2020) suggested that "to provide the *strongest* basis for verifying the prediction of another behavior that has been exposed to an IV, two conditions must be met: (a) the two behaviors must be measured concurrently, and (b) *all* of the *relevant* variables that influence one behavior must have an opportunity to influence the other behavior" (p. 207, emphasis added). It should be noted that if *all* the relevant variables that influence one tier must have an opportunity to influence the other tiers, then the concurrent measurement of behavior is implied because each tier would have to occur at the same time for *all* possible coincidental variables to have the opportunity to affect behavior. Thus, their statement can be simplified to include only part (b) (because part [a] is implied within part [b]).

This suggestion, however, is clearly problematic because it is impossible to ensure that *all* the relevant variables have an opportunity to influence behavior in applied research. Researchers may simply not know all the possible variables that could potentially influence a behavior. Even if they do, it may be impossible to impose that level of control over their participants' lives. This shifts the discussion to determining the *extent* to which experimenters must ensure the most relevant variables have the opportunity to affect the DV across tiers to demonstrate verification. We presume that Cooper et al. (2020) would assert that coincidental variables are relevant variables based on their assertion that concurrent baselines are necessary; however, as discussed earlier, it may be the case that coincidental variables have negligible relevance depending on other aspects of the experiment (e.g., the nature

[2] It is worth emphasizing that this logic is based on affirming the consequent, which is a logical fallacy. Thus, even if NonconMBLs cannot fulfill this requirement, it may not be problematic because fulfilling this requirement should only provide negligible increases in confidence regarding the experimental effect from a perspective based purely on logic.

of the DVs and IVs, other programmed control procedures). We argue that in certain experiments it may be the case that NonconMBLs can still provide enough opportunity for a sufficient number of relevant variables to affect each tier, even if they do not occur concurrently. By permitting a sufficient number of the relevant variables to have an effect, we believe there can be sufficient similarities across tiers to suggest that verification can still be achieved.

Alternative Perspectives

Although in general we agree with Slocum et al.'s (2022) conclusion that ConMBLs and NonconMBLs offer comparable internal validity in many cases and we have argued for even more flexibility in experimental designs in general, it is also important to consider the expectations of investigators and practitioners unfamiliar with behavior analytic research, who may be less familiar with single-subject research designs and the principles of behavior. These investigators and practitioners might not understand the logic supporting the research designs nor the conditions under which more stringent control is (or is not) necessary. For example, it might be more challenging for investigators or practitioners with less familiarity with behavior analytic research to identify what extraneous variables are relevant in a given experiment, making it difficult to identify which design features are necessary to demonstrate experimental control. However, convincing such investigators and practitioners may be necessary to (1) obtain grant funding; (2) disseminate conclusions from specific experiments; (3) have experiments included in meta-analyses and systematic reviews; or (4) establish the credibility of the field of behavior analysis in general. Thus, behavior analysts should also consider the expectations of outside investigators and practitioners.

One example of outside investigators and practitioners delineating requirements for single-subject research designs comes from the What Works Clearinghouse (WWC), which is a federally funded initiative aimed at establishing standards to determine whether experiments should be included as evidence supporting an intervention. The *WWC Standards Handbook* (WWC, 2020) provides an objective rating scale for single-subject research designs and states that tiers within a multiple-baseline design must be concurrent to meet their standards (p. 80). Thus, reviewers for the WWC would not consider any studies using NonconMBLs as evidence supporting an intervention. If other researchers rely on WWC standards when they conduct meta-analyses and systematic reviews, they will exclude all studies using NonconMBLs. In this way, permitting researchers to use NonconMBLs may slow the dissemination of behavior analytic research. Although Slocum et al. (2022) made an excellent argument for the use of NonconMBLs and in general behavior analysts may accept NonconMBLs, researchers should also consider the contingencies in place from the larger scientific community when selecting their research designs. To this end, it may be worth exploring whether the perspective espoused by the WWC is shared by other fields. Further, if other fields do share this perspective, behavior analysts should strive to meet the putatively more rigorous standards of these fields *or* aim to educate others about the nuances of this debate and encourage them to

adjust their standards. Perhaps the best answer would be to continue to use ConMBLs when possible, at least until other fields become more familiar with the merits of NonconMBLs. This latter familiarity may require a presentation of experimental logic, such as those debated in the current series of papers, and dissemination of this logic with other fields such as education, medicine, and allied health professions.

Conclusion

We agree with Slocum et al. (2022) that NonconMBLs and ConMBLs can demonstrate similar amounts of experimental control. That said, it is important to remember that there is a great deal of nuance involved in evaluating whether a specific study demonstrates experimental control. We have argued that researchers and reviewers should avoid making a priori determinations about the relative strengths of different designs, and they should avoid strict adherence to a rigid set of rules when they evaluate experimental designs.

In the end, science is a social endeavor. The purpose of any experiment is to convince other people of a relation between the IV and DV, which includes ruling out plausible alternative explanations of the observed relation within the experiment. Although researchers have developed guidelines for experimental designs to increase the likelihood that an experiment will convince other people of the experimental effect, it should not matter how an experimenter designs their study as long as the experiment provides a convincing demonstration of the relation between the IV and DV.

Compliance with Ethical Standards

Conflict of Interest The authors have no conflicts of interest to declare.

References

Carr, J. E. (2005). Recommendations for reporting multiple-baseline designs across participants. *Behavioral Interventions, 20*(3), 219–224. https://doi.org/10.1002/bin.191

Christ, T. J. (2007). Experimental control and threats to internal validity of concurrent and nonconcurrent multiple baseline designs. *Psychology in the Schools, 44*(5), 451–459. https://doi.org/10.1002/pits.20237

Cooper, J. O., Heron, T. E., & Heward, W. L. (2020). *Applied behavior analysis* (3rd ed.). Pearson Education.

Gast, D. L., Lloyd, B. P., & Ledford, J. R. (2018). Multiple baseline and multiple probe designs. In J. R. Ledford & D. L. Gast (Eds.), *Single case research methodology: Applications in special education and behavioral sciences* (pp. 288–335). Routledge/Taylor & Francis Group. https://doi.org/10.4324/9781315150666

Ghaemmaghami, M., Hanley, G. P., & Jessel, J. (2021). Functional communication training: From efficacy to effectiveness. *Journal of Applied Behavior Analysis, 54*(1), 122–143. https://doi.org/10.1002/jaba.762

Harvey, M. T., May, M. E., & Kennedy, C. H. (2004). Nonconcurrent multiple baseline designs and the evaluation of educational systems. *Journal of Behavioral Education, 13*(4), 267–276. https://doi.org/10.1023/B:JOBE.0000044735.51022.5d

Hayes, S. C. (1981). Single case experimental design and empirical clinical practice. *Journal of Consulting and Clinical Psychology, 49*(2), 193–211. https://doi.org/10.1037/0022-006X.49.2.193.

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single subject research to identify evidence-based practice in special education. *Exceptional Children, 71*(2), 165–179. https://doi.org/10.1177/001440290507100203

Johnston, J. M., Pennypacker, H. S., & Green, G. (2020). *Strategies and tactics of behavioral research and practice* (4th ed.). Routledge/Taylor & Francis Group.

Kazdin, A. E. (2021). *Single-case research designs: Methods for clinical and applied settings* (3rd ed.). Oxford University Press.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2013). Single-case intervention research design standards. *Remedial and Special Education, 34*(1), 26–38. https://doi.org/10.1177/0741932512452794

Lehardy, R. K., Luczynski, K. C., Hood, S. A., & McKeown, C. A. (2021). Remote teaching of publication-quality, single-case graphs in Microsoft Excel. *Journal of Applied Behavior Analysis, 54*(3), 1265–1280. https://doi.org/10.1002/jaba.805

Lloveras, L. A., Tate, S. A., Vollmer, T. R., King, M., Jones, H., & Peters, K. P. (2022). Training behavior analysts to conduct functional analyses using a remote group behavioral skills training package. *Journal of Applied Behavior Analysis, 55*(1), 290–304. https://doi.org/10.1002/jaba.893

McDevitt, M. A., Pisklak, J. M., Dunn, R. M., & Spech, M. (2022). Forced-exposure trials increase suboptimal choice. *Psychonomic Bulletin & Review*. Advance online publication. https://doi.org/10.3758/s13423-022-02092-2

Slocum, T. A., Pinkelman, S. E., Joslyn, P. R., & Nichols, B. (2022). Threats to internal validity in multiple-baseline design variations. *Perspectives on Behavior Science*. https://doi.org/10.1007/s40614-022-00326-1

What Works Clearinghouse. (2020). *What works clearinghouse: Standards handbook, version 4.1*. U.S. Department of Education. https://ies.ed.gov/ncee/wwc/handbooks