Changing criterion designs: Integrating methodological and data analysis recommendations

Rumen Manolov[1 2], Antonio Solanas[1], and Vicenta Sierra[2]

[1] Department of Social Psychology and Quantitative Psychology,

Faculty of Psychology, University of Barcelona (Spain)

[2] Department of Operations, Innovation and Data Sciences,

ESADE Business School, Ramon Llull University (Spain)

Author Note

Correspondence concerning this article should be addressed to Rumen Manolov, Departament de Psicologia Social i Psicologia Quantitativa, Facultat de Psicologia, Universitat de Barcelona, Passeig de la Vall d'Hebron, 171, 08035-Barcelona, Spain. Telephone: +34934031137. E-mail: rrumenov13@ub.edu

Abstract

Changing criterion designs (CCD) are single-case experimental designs that entail a step-by-step approximation to the final level desired for a target behavior. Following a recent review on the desirable methodological features of CCDs, the current text focuses on an analytical challenge: the definition of an objective rule for assessing the correspondence between criterion levels and behavior. We review extensively applied research and methodological recommendations for establishing experimental control, such as the range-bound CCD, and integrate them in a proposal for a formal expression of an objective rule. We illustrate how this proposal can lead to establishing the acceptable level of behavior prior to gathering the intervention phase data and how it leaves space for professional judgement. A discussion of software implementing the suggested proposal is included.

Keywords: changing criterion design, experimental control, overlap, immediacy, variability

Single-case designs entail the repeated measurement of a target behavior in a single unit (individual or group) under the different conditions of the experimentally manipulated independent variable (Horner & Odom, 2014). Criteria for internal and external validity have been suggested (Horner et al., 2005) to improve the quality of single-case designs and for ensuring they can help identifying evidence-based practices (Ganz & Ayres, 2018). The logic of single-case designs and their different types have been presented in textbooks (e.g., Barlow, Nock, & Hersen, 2009; Kennedy, 2005) and journal articles (e.g., Plavnick & Ferreti, 2013).

The current focus is on changing criterion designs (CCDs), in which a brief baseline is followed by an intervention and, during the evaluation of the effect of the intervention, the criterion for reinforcing the behavior of interest is changed (Barlow et al., 2009). In a CCD, there are several sub-phases within the intervention phase, each of them characterized by a different criterion level that has to be met by the target behavior.

CCDs are deemed very useful for increasing or decreasing gradually the level of a behavior, which is a strength for situations in which large immediate changes cannot be expected (Klein et al., 2017). Another strength of CCDs is that they can be applied to non-reversible behaviors (Barlow et al., 2009), unlike ABAB designs and alternating treatments designs. Moreover, demonstrating experimental control via a CCD does not require replication across participants, unlike the most commonly used version of the multiple-baseline design. Furthermore, Ledford and Gast (2014) argue that CCDs are especially useful when the intervention program uses differential reinforcement of other behavior or differential reinforcement of low or high rate of behavior. As a limitation of the CCD, the behavior whose frequency or duration is to be modified should already be in the participant’s repertoire (Ledford & Gast, 2014). Regarding the actual use of CCDs, we can only refer to published research, as reviewed by Shadish and Sullivan

(2011) and Smith (2012), with CCDs representing 2.6% and 4% of the studies, respectively. In absolute terms, the review by Klein et al. (2017) included more than 100 CCD published studies, which suggests that it is worth dedicated efforts to improve the use of CCDs. In relation to the kinds of behaviors studied, according to Klein et al. (2017) the most frequent targets of intervention were academic behaviors, followed by sports/physical behaviors, and self-case behaviors. McDougall, Hawkins, Brady, and Jenkins (2006) present further examples of actual CCD research with children (e.g., increase the number of math problems solved, increase compliance with medication) and adults (e.g., increase work rate, reduce cigarettes smoked).

Aim and Organization of the Article

The current text was prompted by a recent review of CCD studies, presented alongside methodological recommendations (Klein, Houlihan, Vincent, & Panahon, 2017). Moreover, there is scarce literature dedicated specifically to the data analysis of CCDs (except for McDougall, 2005). For instance, Pustejovsky & Ferron (2017) state that most effect sizes work well only with AB, multiple-baseline, and withdrawal [ABAB] designs. In order to build on the work of Klein et al. (2017) and McDougall (2005), we focus on the demonstration of experimental control in a CCD, integrating methodological and data analytical aspects. Specifically, the current text aims to present an objective rule for assessing whether the behavior meets the criterion levels. This objective rule can be understood as a data analytical proposal, addressing one of the features required for demonstrating experimental control.

In terms of the organization of the article, we first refer to how experimental control is demonstrated in CCDs. Second, we review the data patterns indicative of experimental control. Third, we review published CCD research and we focus on how the data have been analyzed.

Fourth, we derive data analysis implications from the review. Fifth, we present our proposal on how to assess objectively whether the data pattern matches the criterion levels sufficiently or not. Finally, a discussion is provided in relation to transparent reporting and publication bias.

Changing Criterion Designs and Experimental Control

"Experimental control" can be considered equivalent to a "functional relation" (Horner & Odom, 2014), established by repetitive experimental manipulations of the independent variable (i.e., the condition) associated with changes in the target behavior (Kennedy, 2005). Another way to understand a functional relation is that the observed data pattern is as predicted according to the characteristics of the design (Riley-Tillman & Burns, 2009). In the context of a CCD, it is possible to establish, via repetitive changes in the criterion for reinforcement, whether there is a causal relation between treatment and behavior (Barlow et al., 2009). Nonetheless, other authors have suggested that a CCD does not provide as strong evidence for a causal effect as withdrawal or multiple-baseline designs (Graham, Karmarker, & Ottenbacher, 2012; Taggart, Taggart, & Siedentop, 1986). Still, it is possible to embed a CCD within a withdrawal design (e.g., Nicol & Hantula, 2001) or within a multiple-baseline design (e.g., Todd, Reid, & Butler-Kisber, 2010). In addition to replicating the basic effect, it is also necessary to further strengthen methodological rigor in relation to, for instance, procedural fidelity (Horner et al., 2005) and interrater observer (Ganz & Ayres, 2018).

Regarding the methodological strategies necessary for controlling for extraneous variables (Horner & Odom, 2014), for a CCD, Hartmann and Hall (1976) recommend that all sub-phases should be long enough to ensure that successive changes in a therapeutic direction are not naturally occurring and that intervention sub-phases should differ in length. Ledford and Gast

(2014) add the importance of differing sub-phase lengths and varying the size of the criterion changes (e.g., not always increasing by 20%). Moreover, the criterion levels are typically pre-selected (Kratochwill et al., 2010) in such a way that the magnitude of change between consecutive criteria is likely to be achievable (Hartmann & Hall, 1976). A priori chosen criterion levels avoid the use of response-guided choices, which may increase the rate of errors when assessing the results (Franklin, Gorman, Beasley, & Allison, 1996), especially false positives.

Another relevant methodological aspect refers to the possibility to revert to a former criterion in order to increase the strength of the evidence of experimental control (Barlow et al., 2009). Specifically, such a reversal may be used as evidence that the change in the behavior is not naturally occurring, for instance, due to maturation (Ledford & Gast, 2014). Klein et al. (2017) make it clear that the mini-reversals to previous criterion levels are not a requirement and according to their review 63% of the studies did not incorporate it.

Finally, regarding the need for replication (Hartmann & Hall, 1976), several suggestions have been made. Kratochwill et al. (2010) state that at least three criterion levels are required for the CCD to meet design standards. Klein et al. (2017) suggest "three or more changes in the criterion level" (p. 53), whereas Ledford and Gast (2014) recommend "four changes in criterion level" (p. 348). Regarding analytical aspects, we consider that for a better assessment of experimental control, an objective rule is necessary to whether the "target behavior changes when and only when the criterion changes and [the] behavior remains at a stated criterion for the duration of the condition" (Ledford, 2018, p. 73). We propose here an objective rule for that purpose, consisting in assessing whether all data points fall into several formally established acceptable ranges, determined before gathering the intervention phase data.

A Review of Data Pattern Recommendations

The article by Hartmann and Hall (1976) is the first and most widely known attempt to present CCDs systematically. These authors refer to several aspects related to the "successful implementation" of a CCD (p. 530). Three of these aspects can be understood as referring to data analysis, as they are related to data features: (a) stable or worsening baseline data; (b) stability after the change in the behavior at one level and before moving to the next criterion level; (c) close correspondence between behavior rate and changes in the criterion is necessary. We could understand that these three data features represent, together with the design requirements regarding phase lengths and the number of treatment phases, evidence for experimental control. Lack of stability has been suggested to be dealt with by returning to a lower criterion (Barlow et al., 2009) or by using statistical analysis (Hartmann & Hall, 1976).

The most frequently mentioned requirement is the "close correspondence" between criterion levels and behavioral measurements (Kazdin, 1989; Kratochwill et al., 2010; Ledford & Gast, 2014), in absence of an objective definition. Additionally, according to Ledford and Gast (2014) the change should be immediate, whereas McDougall et al. (2006) state that the "latency of change" should be "minimal" (p. 3). The basis of the proposal we later make regarding how to define objectively close correspondence is the work of McDougall (2005): the "range-bound CCD", in which a range of acceptable performance is defined. Ledford and Gast (2014) incorporate this option by stating that "the data [should] fall within a pre-established range at each criterion level" (p. 348). The main reason underlying the range-bound CCD is both a clinical one (i.e., ensuring a gradual and potentially healthier change in the behavior) and an analytical one (i.e., offering a consistent but flexible criterion for amount of behavior that conforms to the expected data pattern). Additionally, they propose a quantification called the percentage of conforming data (PCD, representing percentage of measurements within the established limits). Regarding the PCD, the authors do not specify a necessary value, but they do state that it should be used as a supplement and not as a replacement to visual analysis. In relation to CCDs in general and potentially to the range-bound CCD, Ledford and Gast (2014) mention as a limitation the “professional guesswork” (p.349) involved in defining criterion levels and acceptable response ranges. The aim of the current text is to increase the objectivity and reduce the professional guesswork (i.e., decisions on a subjective basis in absence of explicitly spelled out criteria) involved in analyzing CCD data.

Review of Published Research Using CCD

Before proposing an objective tool for data analysis and for the assessment of experimental control, we aim to get familiar with the actual analytical practice in published CCD research.

Search Strategy

Klein et al. (2017) performed a bibliographic search for applied studies using a CCD published before January 2014, identifying 106 studies. We contacted Daniel Houlihan, the corresponding author of the Klein et al. (2017) paper, asking for the references to the 106 studies (see online supplementary material: Appendix A; https://osf.io/aqh3r/). After an extensive search, we were able to locate 101 of these 106 studies (95.28%).

Mirroring the Klein et al. (2017) search strategy in order to include articles published after January 2014, we searched the PsycINFO, ERIC, and PubMed databases using the key words “changing criterion” for articles published in 2014, 2015, 2016, and up to June 2017, plus articles from any period citing the Hartmann and Hall (1976) articles, as per the PsycINFO database. The introduction and discussion section of each newly identified article was examined to locate other studies utilizing a CCD. Inclusion criteria, as used by Klein et al. (2017), were presenting original research data and explicitly stating that a CCD is used. This additional search resulted in locating and retrieving the full text of 27 more articles (see online supplementary material Appendix B; https://osf.io/aqh3r/, also including the list of excluded studies).

Coding

Two aspects were coded. On the one hand, we explored whether there was any pre-defined rule ("mastery criterion") for determining that a criterion level is met and/or for deciding to shift to the next criterion level. We were only looking for explicit sentences in the Method or Results section of the primary studies. For instance, Hartmann and Hall (1976) mention three consecutive days meeting or exceeding the criterion level as such a rule. Additionally, the code "unspecified or several" was assigned to studies using several explicitly stated mastery criteria or to studies not explaining explicitly the criteria used. On the other hand, we categorized the way in which the data are analyzed, as reported in the Method or Results section of the primary studies. The categories used for coding were created on the basis of the six data features from Kratochwill et al. (2010): level, trend, variability, immediacy, overlap, and consistency. We applied these six codes if a report provided either a verbal comment or a quantification. Regarding variability, we applied the code when "stability" was assessed verbally, when the range, the standard deviation, or a range-bound CCD was used, or when individual values were commented. We applied the "level" code when mentioning means and comparing the values to the criterion level established for the sub-phase(s). Regarding immediacy, we applied the code when comments were present on the initial sessions of a sub-phase or when the behavioral change was judged rapid/immediate or not. Regarding consistency, it refers to explicit comments on whether a change in the behavioral level has been replicated consistently, either within the same participant or across participants in the same study. When coding the actual data analysis, we also coded whether the mastery criterion was mentioned again: for instance, commenting the number of sessions required for meeting the criterion.

Results

In terms of the progression in the publication of CCD studies per decades, the results are as follows: 9 studies (7% of the 128 studies included in the review) from 1974 to 1980, 25 studies (19.5%) from 1981 to 1990, 16 studies (12.5%) from 1991 to 2000, 38 studies (29.7%) from 2001 to 2010, and 40 studies (31.3%) from 2011 to 2017.

The results of the use of a mastery criterion are presented in Table 1. Approximately 70% of the studies include a mastery criterion and the most common one (in 28.91% of the studies) was three consecutive sessions meeting or exceeding the criterion level. In terms of any possible changes in time, the following results were obtained regarding the use of a mastery criterion for changing the criterion level: 4 of the 9 studies (44.4%) from 1974 to 1980, 11 of the 25 studies (44%) from 1981 to 1990, 11 of the 16 studies (68.8%) from 1991 to 2000, 28 from 38 studies (73.7%) from 2001 to 2010, and 34 and 40 studies (85%) from 2011 to 2017. Additionally, the number of sessions required until reaching the mastery criterion was explicitly reported as follows: 1 of the 9 studies (11.1%) from 1974 to 1980, 5 of the 25 studies (20%) from 1981 to 1990, 3 of the 16 studies (18.75%) from 1991 to 2000, 7 from 38 studies (18.4%) from 2001 to 2010, and 14 and 40 studies (35%) from 2011 to 2017. Thus, it seems that mastery criteria are becoming increasingly important for determining when to change the criterion level, but the time required for meeting them is not so clearly increasingly used as one of the main outcomes.

The analysis of the CCD data (Table 2) most frequently focuses on commenting verbally or quantifying numerically the level of behavior (in 83.59% of the studies reviewed). Comments on data variability vs. stability and/or quantifications of variability are present in 39.06% of the studies. A reference to the mastery criterion, especially the number of sessions required to meet it, is present in 22.66% of the studies. It should be noted that several studies comment on more than one data feature, which is why the percentages add up to more than 100%.

INSERT TABLES 1 AND 2 ABOUT HERE

Implications Derived from the Recommendations and the Review

The Need for Adequate Quantitative Aids

Visual analysis is expected to focus on several data aspects (Kratochwill et al., 2010) and, therefore, it may provide a holistic assessment, while also being useful for validating numerical results (Parker, Cryer, & Byrns, 2006). Nonetheless, it may be necessary to develop visual aids (Gast & Spriggs, 2014), protocols (Maggin, Briesch, & Chafouleas, 2013), and to train visual analysts to use structured criteria (Wolfe & Slocum, 2015). The reason for looking for complements to visual analysis is that it may not be sufficient to look for dramatic changes (Cooper, Heron, & Heward, 2007), because visual analysis may miss small but clinically meaningful effects (Barnard-Brak, Richman, Little, & Yang, 2018).

In the CCD context, Hartmann and Hall (1976, p. 531) state that “with data characterized by excessive variability or complex trends, a simple visual display may be unconvincing, and statistical procedures may be required”. Therefore, we consider that terms such as “excessive variability” need to be defined operatively given the importance of stability for all single-case designs and for CCD data in particular (Barlow et al., 2009; Hartmann & Hall, 1976; Ledford & Gast, 2014). An objective tool may not only help assessing whether the variability is excessive (e.g., Allen & Kramer, 1990; Figure A1 from online supplementary material: Appendix C; https://osf.io/aqh3r/), but it may also help identifying which behavioral change can be considered as an indicator of intervention effectiveness or experimental control.

Verbal Comments May Not Be Sufficient

Facon, Sahiri, and Rivière (2008) mention a “very close correspondence” (p. 318) between the changes of criterion level and the behavior, but the behavior met or exceeded the criterion levels for only 16 of the 45 measurement occasions and from the graph (Figure A2 from the online supplementary material: Appendix C; https://osf.io/aqh3r/). Using an acceptable range of criterion level plus/minus 10% (as in the example of the range-bound CCD; McDougall et al., 2006), all measurements would actually be included in this range (see upper panel of Figure A2). In contrast, if we apply our proposal for acceptable ranges that do not overlap, the result is not that clear (see lower panel of Figure A2).

Means Do Not Necessarily Provide Useful Information

First, means do not quantify what percentage of the data points meet the criterion level. Second, means do not inform about immediacy: for instance, in Marcus, Sinnott, Bradley and Grey (2010) only means are reported but no emphasis is put on the time needed to reach the mastery criterion of two consecutive sessions with 0 target behavior (see Figure A3 from the online supplementary material: Appendix C; https://osf.io/aqh3r/). Third, means do not inform about stability: for instance, Shaw and Simms (2009, p. 257) state that the intervention was “successful in decreasing challenging behaviors across all students” and only report means, not considering variability (see Figure A4 from Appendix C). Fourth, means do not inform about the amount of overlap with the previous criterion level. For instance, in the Todd et al. (2010) data, the initial values of each criterion level overlap with the previous phase (see Figure A5 from Appendix C), but the authors still conclude that "Lisa and Daniel increased their distance cycled during a 30-min session as a function of the self-regulation intervention" (p. 238). We advocate for a pre-defined acceptable range that would overcome the limitations of reporting mainly means.

Overlap is a Relevant Criterion

McDougall et al. (2006) mention the lack of overlap between the behavioral measures belonging to different sub-phases of the intervention phase as important for demonstrating experimental control. Regarding the studies reviewed, Cannella-Malone, DeBar, and Sigafoos, (2009) include lack of overlap with the baseline data as a sign of progress in the target behavior. Additionally, Ganz and Flores (2008) use the Percentage of nonoverlapping data (PND) as a quantification. However, measures such as the PND and the Nonoverlap of all pairs (NAP) are useful when the data are stable (i.e., without a trend) and do not present outliers, whereas other nonoverlap measures do control for trends (see the review by Parker, Vannest, & Davis, 2011). In relation to our proposal, the acceptable ranges can be established not to overlap across criterion levels and conforming to these ranges would be indicative of lack of excessive variability.

Mastery Criteria and Immediacy May Not Agree Well

The use of mastery criteria can be understood either as a substitute for the close correspondence or as way of defining it objectively. Therefore, not all (100%) data points need to meet the criterion and not from the very first measurement session at the new criterion level. For instance, Schleien, Wehman, and Kiernan (1981) focus on the time needed to reach the mastery criterion of two consecutive completions of the 7 steps (see Figure A6, from the online supplementary material: Appendix C; https://osf.io/aqh3r/, for one of their participants' data). The idea that immediacy is not a strict requirement for demonstrating experimental effectiveness is not well-aligned with Kratochwill et al. (2010) or with the requirements described in McDougall et al. (2006) and in Ledford and Gast (2014). However, accepting lack of immediacy is reflected in the current focus on delayed effects (Levin, Ferron, & Gafurov, 2017; Natesan & Hedges, 2017). In case delayed effects are deemed acceptable, it would be necessary to state explicitly, prior to gathering the data, at which moment the effect is expected to take place. A predefined acceptable range would not require strictly speaking immediately meeting the criterion level; what is required is that the data conform to the range (rather than to its middle line) from the beginning.

Not Any Quantification Is Suitable

The quantification needs to reflect well the features of the CCD. For instance, Edinger, Nelson, Davidson, and Wallace (1978) computed Pearson's correlation between the criterion level for each session and the target behavior value for the same sessions. A correlation only reflects a linear relation between the two sets of numbers, but does not guarantee that the behavior matches the criterion (i.e., that these numbers are similar). Similarly, if we were to quantify the slope across all intervention sub-phases (as proposed by Shadish, Kyse, & Rindskopf, 2013, when discussing applications of multilevel analysis), a steeper slope would not necessarily imply correspondence to the criterion level. Actually, the Pearson correlation and the estimation of slope maybe better related to the idea of using a goal line as a means of assessing whether the behavior is progressing as fast as expected to the final goal (Riley-Tillman & Burns, 2009). Nonetheless, this is not the same procedural aim as in CCDs. Finally, in relation to statistical tests (e.g., Brogan, Falligant, & Rapp, 2017), it is necessary to discuss exactly what information is provided when concluding that the difference between the baseline data and the intervention phases is statistically significant.

It Is Possible to Assess Several Data Features

When analyzing visually single-case designs data, several data features are expected to be assessed (Kratochwill et al., 2010). For CCD published research, we also found instance of such practice. Scott, Scott, and Howe (1998) discuss the immediacy of effect, average level, amount of overlap, and variability quantified via the standard deviation. Similarly, Ganz and Flores (2008) comment on overlap quantified via the PND, means and ranges per criterion level, as well as considering the mastery criterion established. Cervantes and Porretta (2013) used the PCD, compared average levels of behavior across phases and mentioned overlap.

Actual Use of an Acceptable Range

In our review, we found one direct application of the range-bound CCD (Cervantes & Porretta, 2013) and two examples related to this idea. Buisson and Murdock (1995) set that all data points should be included within 10% of the criterion level, whereas Fitterling, Martin, Gramling, Cole, and Milan (1988) set that the last three data points should be within 20% of the criterion level.

Main Proposal and Its Justification

The Proposal in Formulaic Terms

We propose refining the idea for the range-bound CCD by McDougall et al. (2006) by presenting an expression for defining the acceptable range of measurements in such a way that, in case all data conform to the ranges (i.e., PCD=100%), as recommended here, there would a complete nonoverlap between adjacent conditions, which could be represented as NAP equal to 100%.

For the first criterion level (CL_1), the upper limit (UL) and the lower limit (LL) of the acceptable range are defined on the basis of the baseline data: $UL_1 = CL_1 + d_1$ and $LL_1 = CL_1 - d_1$. In this expression, d_1 is the half-width of the first interval and it is defined as $d_1 = scatter_A \times p_1; 0 < p_1 < 1$, where $scatter_A$ is some measure of baseline variability, expressed in the same units of the variable of interest. In single-case research, the range is the usual measure of variability, preferred over the standard deviation (Ledford, Lane, & Severini, 2018), which was mirrored in our review and, specifically, several studies explicitly comment on the range of the baseline data, before describing the results from the intervention phase (e.g., Carson, Moosa, Theurer, & Cardy, 2012). If we use range as a measure of variability, $scatter_1 = \max(y_{Ai}) - \min(y_{Ai})\,; i = 1, 2, \ldots, n_A$, where n_A is the number of baseline phase measurements. In case there are clear outliers[1] in the baseline data, a resistant measure such as the median of the absolute deviations from the median (MAD) can be used. The other term in the expression is p_1, representing the proportion of the baseline phase scatter measure that is used when defining the acceptable range for the first criterion level. On the one hand, the greater the value of p_1, the wider and less stringent the acceptable range. On the other hand, a smaller value of p_1 can compensate for a measure of scatter, affected by a potential outlier in the baseline data.

For subsequent criterion levels ($j = 2, 3, \ldots, k$), the acceptable range is defined as follows: $UL_j = CL_j + d_j$ and $LL_j = CL_j - d_j$. In these latter two expressions, the half-width of the interval (d_j) is computed on the basis of the difference between the current criterion level (CL_j)

[1] Outliers can be detected quantitatively using the rule incorporated in the boxplot or using a rule based on MAD itself (Leys, Ley, Klein, Bernard, & Licata, 2013). However, it is not clear whether such rules are useful when few data points are available (like in a typical baseline from a CCD). Therefore, it may be more reasonable to define what an outlier is on the basis of substantive knowledge. In absence of such knowledge, we recommend using the range instead of MAD, as the range is the more commonly used and more easily visually represented.

and the previous criterion level (CL_{j-1}), while also taking into account the half-width of the previous criterion level (d_{j-1}): $d_j = \left(\left|CL_j - CL_{j-1}\right| - d_{j-1}\right) \times p_j; 0 < p_j < 1$.

In the expression for d_j, the term p_j represents a proportion and enables making the acceptable range narrower and more stringent (for smaller values of p_j) or wider and less stringent (for greater values of p_j). Actually, the restriction $0 < p_j < 1$ is introduced, given that in case $p_j = 0$ the acceptable range would be represented only by the criterion level value (CL_j), whereas in case $p_j = 1$ the acceptable ranges of two adjacent criterion levels would just overlap.

The Proposal and the Choice of p_1 and p_j via an Example

The main questions that arises from the formal presentation of the proposal is how to choose p_1 and the subsequent p_j values. The initial answer is by trial and error, in such a way as to produce acceptable ranges that seem reasonable to the researcher. Given that this answer entails subjectivity, it is crucial that the acceptable ranges are chosen before gathering and representing graphically the intervention phase data. Actually, in the following example, we have intentionally excluded the intervention phase data in order to show that they are not necessary for defining the criterion levels or the acceptable ranges. Next, we illustrate how the choice of p_1 and p_j values affects the ranges defined and what ranges could be considered “reasonable” in general, although the researcher’s criterion would be useful to fine-tune the acceptable ranges.

The example represented in Figures 1 and 2 includes two criterion levels: 6.7 and 10.2, as retrieved from the graphical representation from Todd et al. (2010) for a participant called Lisa. The first sub-phase consists of 6 sessions and the second of 7. The acceptable range around the first criterion level is decided on the basis of the baseline variability $scatter_A = \max(y_{Ai}) -$

$\min(y_{Ai}) = 6 - 2$. For the sake of illustration, the first proportion is chosen to be $p_1 = 0.5$, because it would lead to a half-width of the first acceptable range equal to half the baseline data range (i.e., the first acceptable range would be as large as the baseline range). However, a different value can be chosen by trial and error, as illustrated in Figure 1. Note that a value of $p_j = 0.5, j \neq 1$, would mean that the half-width of the acceptable range would be half the distance between the criterion level j and, for instance, the upper bound of the acceptable range for the previous criterion level.

For the top left panel of Figure 1, $d_1 = scatter_A \times p_1 = 4 \times 0.5 = 2$ and, therefore, the upper limit for the first criterion level is $6.7 + 2 = 8.7$ and the lower limit is $6.7 - 2 = 4.7$. For the second criterion level, the initial proportion chosen is also $p_2 = 0.5$ and the half-width of the interval is computed as $d_j = \left(\left|CL_j - CL_{j-1}\right| - d_{j-1}\right) \times p_j = (|10.2 - 6.7| - 2) \times 0.5 = (3.5 - 2) \times 0.5 = 0.75$. The upper limit for the second criterion level is $10.2 + 0.75 = 10.95$ and the lower limit is $10.2 - 0.75 = 9.45$. Note that the same value of p_j does not entail the same half-width, because the half-width also depends on the difference between adjacent criterion levels. Specifically, for the same p_j, if the difference between adjacent criterion levels is smaller, the acceptable range would be narrower. In summary, p_1 is the proportion of baseline variability to be used when defining the first acceptable range and p_2 is the proportion of the distance between the upper limit of the acceptable range from the first sub-phase and the criterion level of the second sub-phase.

Regarding the effect of different values of p_j, from Figure 1 it is clear how larger values of p_1 entail a wider acceptable range for the first sub-phase and smaller values entail a narrower range. When the first acceptable range is narrow and the two adjacent criterion levels are clearly separated (top middle and top right panels of Figure 1), then the positive relation between the value of p_2 and the width of the acceptable range is also clear. When the first acceptable range is wide (bottom left and bottom middle panels of Figure 1), the subsequent second acceptable range has to be narrower in order to avoid overlap between the acceptable ranges. In such case, the value of p_2 is also related positively to the width of interval the acceptable range. Nevertheless, it has to be noted that p_1 and p_2 do not represent proportions of the corresponding criterion levels, as the bottom left panel of Figure 1 illustrates the fact that p_2 being greater than p_1 does not imply that the acceptable range for the second sub-phase is wider.

Regarding the a priori choice of "reasonable" acceptable ranges, one option is to establish more restrictive ranges for later sub-phases as compared to initial ones. This option reflects well the possibility that the final aim is elimination of a problematic behavior (i.e., a required percentage of occurrence of 0%) or the complete instauration of a desirable behavior (i.e., a required percentage of occurrence of 100%). Accordingly, when the target behavior approaches the floor and ceiling effects, it is logical that the variability ought to be lower. In the current example, this can be achieved with several different combinations of p_1 and p_2, as illustrated in Figure 1. Which is better is up to the researcher's criterion. Another option for establishing "reasonable" acceptable ranges is to make their width approximately the same in absolute terms. This is what we tried to achieve in the bottom right panel of Figure 1 and later used in Figure 2: we chose relatively wide acceptable ranges due to the variable baseline data.

Overall, it is not clear whether the acceptable ranges need to have the same width in absolute terms, the same relative width in relation to the criterion level, or these ranges should be progressively narrower. For instance, the example from McDougall (2013) includes several ranges of equal relative width (10% of the criterion level), but the absolute width is larger for higher criterion levels (established for later sub-phases), leading to a greater range of acceptable values as the behavior approaches the final goal. In any case, the expressions presented here make possible any of these options.

Regarding the application of the acceptable range as an objective rule, as per the lower panel of Figure 2, we could say that not all data conform to the acceptable ranges: 2 values from each sub-phase fail to do so. If we applied to range-bound CCD on the basis of the criterion level plus/minus 10% (Cervantes & Porretta, 2013; McDougall et al., 2006), the ranges are actually smaller for these data (unlike for Figure A2) and thus it is even clearer that there is not a close correspondence between the criterion level and the behavior observed. In contrast, recall that the authors' conclusion was that "Lisa and Daniel increased their distance cycled during a 30-min session as a function of the self-regulation intervention" (Todd et al., 2010, p. 238). In absence of close correspondence (Figure 2) and more sub-phases, it is not possible to infer causality.

INSERT FIGURES 1 AND 2 ABOUT HERE

Justification

First, the proposal provides an objective operative definition for ambiguous phrases such as close correspondence between behavior and criterion level, reducing the amount of guesswork (Ledford & Gast, 2014) involved in defining acceptable response ranges, although researchers still can incorporate their judgment in the definition of the proportions p_j. Objective ranges drawn on the same graph would reduce the potential deleterious effect of inappropriately constructed graphs (Dart & Radley, 2018). Second, it allows dealing with several data aspects simultaneously: (a) variability of the baseline data represented in the term $scatter_A$; (b) variability in the intervention phase data as compared to the acceptable ranges; (c) overlap: in case the data conform to the acceptable ranges, there would be no overlap across criterion levels; in that sense, the change in level would be clear; (d) immediacy of the effect: there is certain tolerance for not meeting or exceeding the criterion level immediately, but the data should conform to the acceptable range immediately; and (e) the mastery criterion can be defined as all values falling within the pre-defined acceptable range for the whole duration of the criterion level or can be used as a supplement to the acceptable range. Third, the expressions presented here have been incorporated in a web-based Shiny application (https://manolov.shinyapps.io/ChangingCriterion; see online supplementary material (Appendix D; https://osf.io/aqh3r/) of the input and numerical and graphical output.

In terms of statistical alternatives, Bayesian analysis has recently illustrated by Aerts (2015) for a CCD, but its use is limited to estimating intercepts and slopes and does not deal with all data features mentioned in the previous paragraph, which are also the typical ones on which visual analysis focuses (McDougall, 2013).

Use of the Proposal

In order to ensure experimental control, we suggest that the PCD should be 100%, which would also ensure that there is no overlap. We consider that this analytical option has to be used jointly with the methodological recommendations reviewed in the introduction (e.g., varying phase lengths for the different criterion levels, replicating the findings across several participants or behaviors), because it is only one piece of information that needs to be integrated with the remaining evidence.. In addition to assessing experimental control, it has to be assessed whether the results are clinically meaningful and socially valid (Horner et al., 2005).

Discussion

The Acceptable Range and Professional Judgment

We integrated the pieces of evidence on experimental control with the work of McDougall et al. (2006) and translated them into an objective definition of the acceptable range of behavior under the different criterion levels. Using an acceptable range would be a step forward in overcoming the insufficient reporting present when using visual analysis (Ledford et al., 2018) and it is well-aligned with McDougall's (2013) aim to make evaluating intervention efficacy easier.

Professional knowledge can be used when defining the values of the proportions $p_1$and p_j for establishing the acceptable ranges and when choosing the criterion levels, prior to gathering the data or after seeing the performance at the current criterion level. The use of a mastery criterion is also possible. However, the determination of sub-phase length on the basis of when the mastery criterion is achieved and the determination of the next criterion level on the basis of the performance observed at the current criterion level may lead to interpreting *any* data pattern as showing experimental control. Similarly, Hwang, Levin, and Johnson (2018) claim that the moment of introduction of the intervention should be determined, either systematically or at random, before the study, in order to reduce the possibility of introducing experimenter bias.

Reporting Minima

Our review of published CCD studies indicated a great variety in the way in which the results are commented and presented, except for the almost universal graphical representation. Thus, some systematization of reporting would be very useful. Several pieces of information need to be provided in order to boost transparent reporting: (a) whether criterion levels and sub-phase lengths for each criterion level were determined prior to obtaining the (intervention) measurements or the moments of change in criterion level depended on the data meeting a certain pattern; (b) whether the effect is expected to be immediate or not and how immediacy will be assessed; (c) whether data overlap across criterion levels is considered acceptable or not and how it will be assessed, and (d) whether there is any criterion for assessing data stability vs. variability. In terms of the assessment of intervention effectiveness, besides discussing the presence of a functional relation and the magnitude of effect, any assessment of social validity has to be reported as well (Horner et al., 2005), given that it is relevant for judging whether the intervention is useful and feasible.

The Need for Experimental Control and Publication Bias

Experimental control has been deemed an indispensable basis for inferring a causal relation between the intervention and the target behavior (Kratochwill et al., 2010). However, Shadish, Zelinsky, Vevea and Kratochwill (2016) state that single-case designs "should not require a functional relation any more than a randomized experiment should be required to have a large effect size" (p. 668). Thus, in order to avoid misrepresentation of the evidence due to selective submission or acceptance for publication only oif studies demonstrating experimental control, we concur with Tincani and Travers (2018) that all available findings stemming from studies meeting design standards is valuable. In that sense, good practice CCDs (as per Klein et al., 2017) should be sufficient for reporting and publishing the study, regardless of its results.

Limitations and Future Research

The proposed objective rule cannot be used as the sole indicator of experimental control, as the design and procedural features allowing for controlling confounding variables are essential. One line of future research is related to the requirement for varying sub-phase lengths for the different criterion levels (Ledford & Gast, 2014). One option is to use randomization. For instance, in a

CCD with four criterion levels, the sub-phase lengths could be selected at random from a series of possibilities 3-4-5-6, 3-4-6-5, 3-5-4-6, …, up to 6-5-4-3. There would be $4! = 4 \times 3 \times 2 \times 1 = 24$ possible sets of sub-phase lengths and one of them will be selected. Afterwards, a randomization test (Edgington & Onghena, 2007) could be used on the basis of the randomization performed. The specific test statistic to be used can be the PCD, which would be more in-line with the CCD logic than a mean difference between adjacent conditions. A simulation study can be performed to study the Type I error rates and the statistical power.

A second line for future discussion refers to the need for baseline stability understood by Hartmann and Hall (1976) as zero slope or changing in a counter-therapeutic direction. However, fitting a trend line to a short baseline may not be reliable. Other to-be-discussed options include: (a) the two-standard-deviations band rule (Pfadt & Wheeler, 1995) based on an assumption of normality; (b) the stability envelope (Gast & Spriggs, 2014: 80% of the data falling within the median plus/minus 25% of it); (c) Buisson and Murdock's (1995) rule that each baseline measurement should not vary more than 50% from the baseline mean; and (d) the Easterbrooks and Stoner's (2006) example of requiring at least ten observation points if stability is not reached, considering that such baseline lengths are present in only 15% of the published CCD studies, according to our review. In absence of a statistically sound definition of stability, incorporating a measure of baseline variability in the definition of the acceptable range for the first criterion level (as proposed here) seems to be a justified approach.

Finally, another relevant line of research refers to the meta-analytical integration of the results of studies using different single-case designs. We envision several options for quantitative integration. One option is to count, for each study, whether there is strong evidence (100% PCD as applied to the acceptable ranges) or moderate evidence for a functional relation. The same assessment could be performed for other single-case designs, for instance requiring a ratio of at least 3:1 comparisons with effect per each comparison without effect (Maggin et al., 2013). Afterwards, a vote-counting procedure can be used to integrate the results of several studies (Bushman & Wang, 2009) or the amount of evidence available can be compared to the 5-3-20 rule (Kratochwill et al., 2013). A second option could be based on mean difference, avoiding the need for dichotomizing the results (strong evidence or not). In a CCD, it is possible to compare the baseline condition to the last intervention sub-phase (Faith, Allison, & Gorman, 1996; e.g., Noles, Epstein, & Jones, 1976) to compare pairs of adjacent phases (Shadish et al., 2013; e.g., Nicol & Hantula, 2001) and average these values for the whole CCD. A standardized mean difference can be computed for other single-case designs as well (Odom, Barton, Reichow, Swaminathan, & Pustejovsky, 2018). We consider that counting the studies with evidence and quantifying the magnitude of effect can complement each other. These two strategies for quantitative integration should not be considered as alternatives, because the magnitude of difference between criterion levels or between sub-phase means of measurements does not inform about whether the observed data pattern matched the criterion levels. A third option for quantitative integration is to combine NAP values, which quantify overlap[2]. An average NAP can be computed from the NAP values of adjacent (sub)phases and, afterwards, a weighted average across studies can be obtained, for instance, weighting by the number of participants in a study (e.g., Burns, Zaslofsky, Kanive, & Parker, 2012).

[2] Wolery, Busick, Reichow, and Barton (2010) state that overlap measures do not quantify magnitude of effect. Carter (2013) disagrees, although he does recognize that a ceiling effect when complete nonoverlap is achieved may hinder distinguishing between different sizes of effect. The interested readers may consult Kelley and Preacher (2012) for a thorough discussion of effect size and decide for themselves.

References

Aerts, X. Q. (2015). *Time series data analysis of single subject experimental designs using Bayesian estimation*. Doctoral dissertation. Retrieved October 13, 2017 from https://digital.library.unt.edu/ark:/67531/metadc804882/m2/1/high_res_d/dissertation.pdf

Allen, S. J., & Kramer, J. J. (1990). Modification of personal hygiene and grooming behaviors with contingency contracting: A brief review and case study. *Psychology in the Schools, 27*, 244–251.

Barlow, D., Nock, M., & Hersen, M. (2009). *Single case experimental designs: Strategies for studying behavior change* (3rd Ed.). Boston, MA: Allyn and Bacon.

Barnard-Brak, L., Richman, D. M., Little, T. D., & Yang, Z. (2018). Development of an in-vivo metric to aid visual inspection of single-case design data: Do we need to run more sessions? *Behaviour Research and Therapy, 102*, 8-15.

Brogan, K. M., Falligant, J. M., & Rapp, J. T. (2017). Interdependent group contingencies decrease adolescents' disruptive behaviors during group therapy: A practitioner's demonstration. *Behavior Modification, 41*, 405–421.

Buisson, G. J., & Murdock, J. Y. (1995). Effects of tokens on response latency of students with hearing impairments in a resource room. *Education and Treatment of Children, 18*, 408–421.

Burns, M. K., Zaslofsky, A. F., Kanive, R., & Parker, D. C. (2012). Meta-analysis of incremental rehearsal using phi coefficients to compare single-case and group designs. *Journal of Behavioral Education, 21*, 185-202.

Bushman, B. J., & Wang, M. C. (2009). Vote-counting procedures in meta-analysis. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed.). New York: Russell Sage Foundation.

Cannella-Malone, H. I., DeBar, R. M., & Sigafoos, J. (2009). An examination of preference for augmentative and alternative communication devices with two boys with significant intellectual disabilities. *Augmentative Alternative Communication*, *25*, 262–273.

Carson, L., Moosa, T., Theurer, J., & Cardy, J. O. (2012). The collateral effects of PECS training on speech development in children with autism. *Canadian Journal of Speech-Language Pathology and Audiology, 36*, 182–195.

Carter, M. (2013). Reconsidering overlap-based measures for quantitative synthesis of single-subject data what they tell us and what they don't. *Behavior Modification*, *37,* 378–390.

Cervantes, C. M., & Porretta, D. L. (2013). Impact of after school programming on physical activity among adolescents with visual impairments. *Adapted Physical Activity Quarterly, 30*, 127–146.

Cooper, J.O., Heron, T.E., & Heward, W.L. (2007). *Applied behavior analysis* (2nd ed.). Upper Saddle River, NJ: Pearson Education.

Dart, E. H., & Radley, K. C. (2018). Toward a standard assembly of linear graphs. *School Psychology Quarterly, 33*, 350-355.

Easterbrooks, S. R., & Stoner, M. (2006). Using a visual tool to increase adjectives in the written language of students who are deaf or hard of hearing. *Communication Disorders Quarterly, 27*, 95–123.

Edgington, E. S., & Onghena, P. (2007). *Randomization tests* (4th ed.). London, UK: Chapman & Hall/CRC.

Edinger, J. D., Nelson, W. M., Davidson, K. M., & Wallace, J. (1978). Modification of smoking behaviors in a correctional institution. *Journal of Clinical Psychology, 34*, 991–998.

Facon, B., Sahiri, S., & Rivière, V. (2008). A controlled single-case treatment of severe long-term selective mutism in a child with mental retardation. *Behavior Therapy, 39,* 313–321.

Faith, M. S., Allison, D. B., & Gorman, D. B. (1996). Meta-analysis of single-case research. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 245-277). Mahwah, NJ: Lawrence Erlbaum Associates.

Fitterling, J. M., Martin, J. E., Gramling, S., Cole, P., & Milan, M. A. (1988). Behavioral management of exercise training in vascular headache patients: An investigation of exercise adherence and headache activity. *Journal of Applied Behavior Analysis, 21*, 9–19.

Franklin, R. D., Gorman, B. S., Beasley, T. M., & Allison, D. B. (1996). Graphical display and visual analysis. In R. D. Franklin, D. B. Allison & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 119–158). Mahwah, NJ: Lawrence Erlbaum.

Ganz, J. B., & Ayres, K. M. (2018). Methodological standards in single-case experimental design: Raising the bar. *Research in Developmental Disabilities, 79*, 3-9.

Ganz, J. B., & Flores, M. M. (2008). Effects of the use of visual strategies in play groups for children with autism spectrum disorders and their peers. *Journal of Autism and Developmental Disorders, 38*, 926–940.

Gast, D. L. & Spriggs, A. D. (2014). Visual analysis of graphic data. In D. L. Gast & J. R. Ledford (Eds.), *Single case research methodology: Applications in special education and behavioral sciences* (pp. 176–210). London, UK: Routledge.

Graham, J. E., Karmarkar, A. M., & Ottenbacher, K. J. (2012). Small sample research designs for evidence-based rehabilitation: Issues and Methods. *Archives of Physical Medicine and Rehabilitation, 93*, S111-S116.

Hartmann, D. P., & Hall, R. V. (1976). The changing criterion design. *Journal of Applied Behavior Analysis, 9*, 527–532.

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*, 165-179.

Horner, R. J., & Odom, S. L. (2014). Constructing single-case research designs: Logic and options. In T. R. Kratochwill & J. R. Levind (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 27-51). Washington, DC: American Psychological Association.

Hwang, Y., Levin, J. R., & Johnson, E. W. (2018). Pictorial mnemonic-strategy interventions for children with special needs: Illustration of a multiply randomized single-case crossover design. *Developmental Neurorehabilitation*, *21*, 223-237.

Kazdin, A. E. (1989). *Behavior modification in applied settings* (4th ed.). Pacific Grove, CA: Brooks/Cole Publishing Company.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods, 17*, 137–152.

Kennedy, C. H. (2005). *Single-case designs for educational research*. Boston, MA: Pearson.

Klein, L. A., Houlihan, D., Vincent, J. L., & Panahon, C. J. (2017). Best practices in utilizing the changing criterion design. *Behavior Analysis in Practice, 10*, 52–61.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M. & Shadish, W. R. (2010). *Single-case designs technical documentation*. Retrieved from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/Docs/ReferenceResources/wwc_scd.pdf

Kratochwill, T. R., Hitchcock, J. H., 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*, 26-38.

Ledford, J. R. (2018). No randomization? No problem: Experimental control and random assignment in single case research. *American Journal of Evaluation*, *39*, 71-90.

Ledford, J. R., & Gast, D. L. (2014). Combination and other designs. In D. L. Gast & J. R. Ledford (Eds.), *Single case research methodology: Applications in special education and behavioral sciences* (pp. 346–377). London, UK: Routledge.

Ledford, J. R., Lane, J. D., & Severini, K. E. (2018). Systematic use of visual analysis for assessing outcomes in single case design studies. *Brain Impairment, 19*, 4-17.

Levin, J. R., Ferron, J. M., & Gafurov, B. S. (2017). Additional comparisons of randomization-test procedures for single-case multiple-baseline designs: Alternative effect types. *Journal of School Psychology*, *63*, 13-34.

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology, 49*, 764-766.

Maggin, D. M., Briesch, A. M., & Chafouleas, S. M. (2013). An application of the What Works Clearinghouse standards for evaluating single-subject research: Synthesis of the self-management literature base. *Remedial and Special Education, 34*, 44-58.

Marcus, A., Sinnott, B., Bradley, S., & Grey, I. (2010). Treatment of idiopathic toe-walking in children with Autism using GaitSpot Auditory Speakers and simplified habit reversal. *Research in Autism Spectrum Disorders, 4*, 260–267.

McDougall, D. (2005). The range-bound changing criterion design. *Behavioral Interventions, 20*, 129-137.

McDougall, D. (2013). Applying single-case design innovations to research in sport and exercise psychology. *Journal of Applied Sport Psychology, 25*, 33–45.

McDougall, D., Hawkins, J., Brady, M., & Jenkins, A. (2006). Recent innovations in the changing criterion design: Implications for research and practice in special education. *The Journal of Special Education, 40*, 2–15.

Natesan, P., & Hedges, L. V. (2017). Bayesian unknown change-point models to investigate immediacy in single case designs. *Psychological Methods, 22*, 743-759.

Nicol, N., & Hantula, D. A. (2001). Decreasing delivery drivers' departure times. *Journal of Organizational Behavior Management, 21*, 105–116.

Noles, K. O., Epstein, L. H., & Jones, H. T. (1976). Instruction control of sleep reduction. *Perceptual and Motor Skills, 42*, 669–670.

Odom, S. L., Barton, E. E., Reichow, B., Swaminathan, H., & Pustejovsky, J. E. (2018). Between-case standardized effect size analysis of single case designs: Examination of the two methods. *Research in Developmental Disabilities, 79,* 88-96.

Parker, R. I., Cryer, J., & Byrns, G. (2006). Controlling baseline trend in single-case research. *School Psychology Quarterly, 21*, 418-443.

Parker, R. I., Vannest, K. J., & Davis, J. L. (2011). Effect size in single-case research: A review of nine nonoverlap techniques. *Behavior Modification, 35*, 303-322.

Pfadt, A., & Wheeler, D. J. (1995). Using statistical process control to make data-based clinical decisions. *Journal of Applied Behavior Analysis*, *28*, 349–370.

Plavnick, J. B., & Ferreri, S. J. (2013). Single-case experimental designs in educational research: A methodology for causal analyses in teaching and learning. *Educational Psychology Review*, *25*, 549-569.

Pustejovsky, J. E., & Ferron, J. M. (2017). Research synthesis and meta-analysis of single-case designs. In J. M. Kauffman, D. P. Hallahan, & P. C. Pullen (Eds.), *Handbook of special education* (2nd ed.) (pp. 168-186). New York, NY: Routledge.

Riley-Tillman, T. C., & Burns, M. K. (2009). *Evaluating educational interventions: Single-case design for measuring response to intervention*. London, UK: The Guilford Press.

Schleien, S. J., Wehman, P., & Kiernan, J. (1981). Teaching leisure skill to severely handicapped adults: An age-appropriate darts game. *Journal of Applied Behavior Analysis, 14*, 513–519.

Scott, D., Scott, L. M., & Howe, B. L. (1998). Training anticipation for intermediate tennis players. *Behavior Modification, 22*, 243–261.

Shadish, W. R., Kyse, E. N., & Rindskopf, D. M. (2013). Analyzing data from single-case designs using multilevel models: New applications and some agenda items for future research. *Psychological Methods, 18*, 385-405.

Shadish, W. R., Zelinsky, N. A. M., Vevea, J. L., & Kratochwill, T. R. (2016). A survey of publication practices of single-case design researchers when treatments have small or large effects. *Journal of Applied Behavior Analysis, 49*, 1–18.

Taggart, A. C., Taggart, J., & Siedentop, D. (1986). Effects of a home-based activity program: A study with low fitness elementary school children. *Behavior Modification, 10*, 487–507.

Tincani, M., & Travers, J. (2018). Publishing single-case research design studies that do not demonstrate experimental control. *Remedial and Special Education*, *39*, 118–128.

Todd, T., Reid, G., & Butler-Kisber, L. (2010). Cycling for students with ASD: Self-regulation promotes sustained physical activity. *Adapted Physical Activity Quarterly, 27*, 226–241.

Wolery, M., Busick, M., Reichow, B., & Barton, E. E. (2010). Comparison of overlap methods for quantitatively synthesizing single-subject data. *Journal of Special Education, 44*, 18–29.

Wolfe, K., & Slocum, T. A. (2015). A comparison of two approaches to training visual analysis of AB graphs. *Journal of Applied Behavior Analysis, 48*, 472–477.

Table 1

Mastery criteria used in the published studies reviewed

Mastery criterion	Absolute frequency	Percentage
No mastery criterion explicitly used	39	30.47
1 session	13	10.16
2 consecutive sessions	14	10.94
3 consecutive sessions	37	28.91
4 or more consecutive sessions	10	7.81
2 out of 3 consecutive sessions	4	3.13
3 out of 4 consecutive sessions	1	0.78
2 non-consecutive sessions	2	1.56
3 non-consecutive sessions	2	1.56
4 or more non-consecutive sessions	2	1.56
unspecified or several	4	3.13

Table 2

Data features on which the analysis (i.e., Results section) of the studies reviewed focus. The sum of absolute frequencies is greater than 128 and the sum of percentages is greater than 100, because most studies commented on more than one data feature.

Data features	Absolute frequency	Percentage of studies
level (incl. comparison to criterion level)	107	83.59
variability	50	39.06
trend	7	5.47
overlap	7	5.47
immediacy	24	18.75
consistency in replication within or across participants	36	28.13
mastery criterion (if met or time to meet)	29	22.66

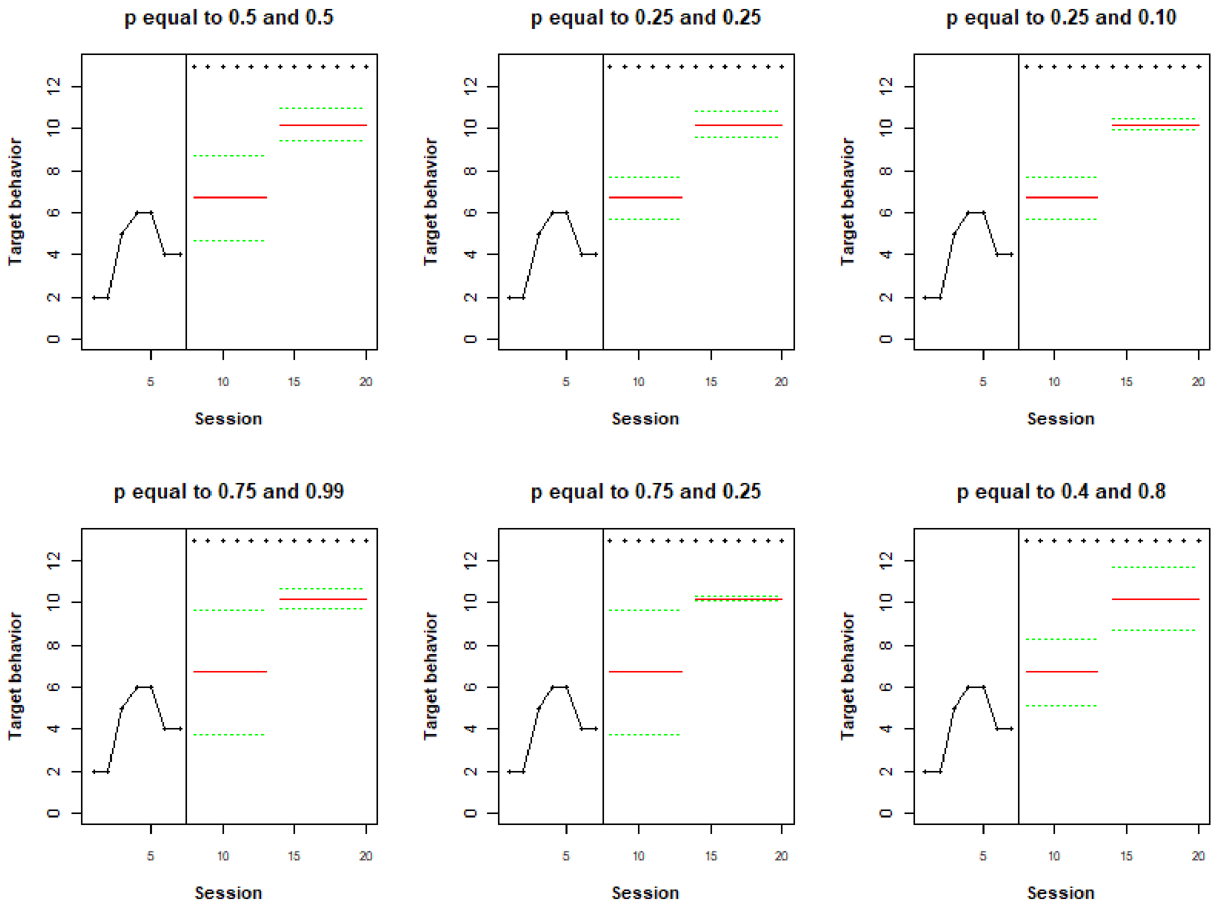


Figure 1. Example of defining acceptable ranges of behavior across criterion levels, using the data by Todd, Reid, and Butler-Kisber (2010) on distance cycled, in number of circuits completed per session, by participants called Lisa. The solid horizontal lines represent the criterion levels, whereas the dotted horizontal lines around them represent the upper and lower limits for several combinations of values of p_1 and p_2, for the two intervention sub-phases.

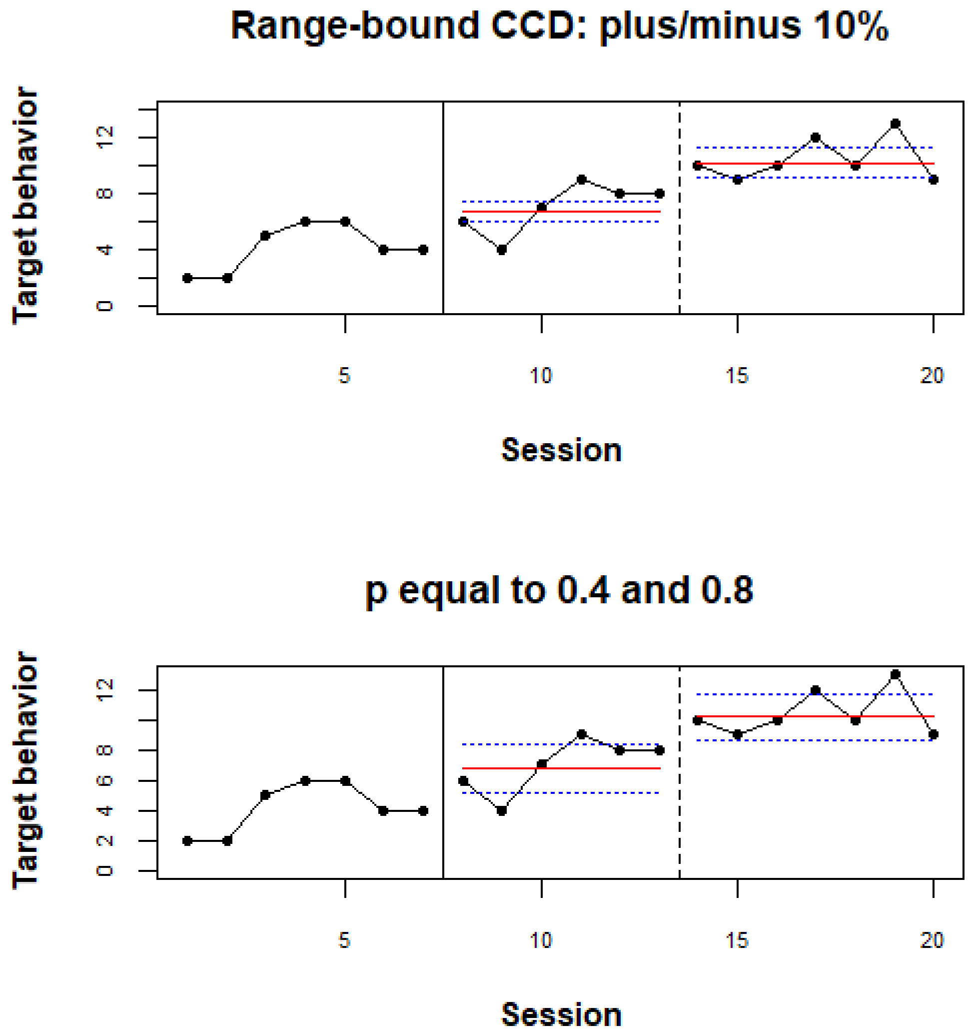


Figure 2. Application of the range-bound definition by McDougall (2005) and the proposal for defining acceptable ranges (the data are the same as for Figures 5 and 7).

Appendix A: References to the Empirical Articles Reviewed by Klein et al. (2017)

Marked with asterisk are the studies included in the current review:

*Alberto, P.A., Waugh, R. E., Fredrick, L. D., & Davis, D. H. (2013). Sight word literacy: a functional-based approach for identification and comprehension of individual words and connected text. *Education and Training in Autism and Developmental Disabilities, 48*, 332–350.

*Allen, K. D., Bernstein, B., & Chait, D. H. (1991). EMG biofeedback treatment of pediatric hyperfunctional dysphonia. *Journal of Behavior Therapy and Experimental Psychiatry, 22*, 97–101.

*Allen, K. D., & Evans, J. H. (2001). Exposure-based treatment to control excessive blood glucose monitoring. *Journal of Applied Behavior Analysis, 34*, 497–500.

*Allen, S. J., & Kramer, J. J. (1990). Modification of personal hygiene and grooming behaviors with contingency contracting: A brief review and case study. *Psychology in the Schools, 27*, 244–251.

*Barahona, C., DuBard, M., Luiselli, J. K., & Kesterson, J. (2013). School-based feeding intervention to increase variety and quantity of foods consumed by an adolescent with Autism. *Clinical Practice in Pediatric Psychology, 1*, 361–368.

*Bates, P., Renzaglia, A., Clees, T. (1980). Improving the work performance of severely/profoundly retarded young adults: The use of a changing criterion procedural design. *Education and Training of the Mentally Retarded, 15*, 95–104.

*Belles, D., & Bradlyn, A. S. (1987). The use of the changing criterion design in achieving controlled smoking in a heavy smoker: A controlled case study. *Journal of Behavior Therapy and Experimental Psychiatry*, *18*, 77–82.

*Bernard, M. E., Dennehy, S., & Keefauver, L. W. (1981). Behavioral treatment of excessive coffee and tea drinking: A case study and partial replication. *Behavior Therapy, 12*, 543–548.

*Bigelow, K. M., Huyen, K. M., & Lutzker, J. R. (1993). Using a changing criterion design to teach fire escape to a child with developmental disabilities. *Journal of Developmental and Physical Disabilities, 5*, 121–128.

*Birkan, B., Krantz, P. J., & McClannahan, L. E. (2011). Teaching children with autism spectrum disorders to cooperate with injections. *Research in Autism Spectrum Disorders, 5*, 941–948.

Brooks, P. (1995). A comparison of the effectiveness of different teaching strategies in teaching spelling to a student with severe specific learning difficulties/dyslexia. *Educational and Child Psychology, 12*, 80–88.

Brown, R. E., Copeland, R. E., & Hall, R. V. (1974). School phobia: Effects of behavior modification treatment applied by an elementary school principal. *Child Study Journal, 4*, 125–133.

Bryan, C., & Dickie, R. F. (1976). The use of a changing criterion design in the elimination of talk-out behavior. *School Application of Learning Theory, 8*, 32–37.

*Buckingham, H., McLaughlin, T. F., & Hunsaker, D. (1978). Increasing oral responses in a special education student with a token program. *Education and Treatment of Children, 1*, 19–24.

*Buisson, G. J., & Murdock, J. Y. (1995). Effects of tokens on response latency of students with hearing impairments in a resource room. *Education and Treatment of Children, 18*, 408–421.

*Burkwist, B. J., Mabee, W. S., & McLaughlin, T. F. (1987). The effect of a daily report card system on inappropriate classroom verbalization with a junior high school learning-disabled student. *Techniques, 3*, 265–272.

*Burley, S., Gutkin, T. B., & Wendy, N. (1994). Assessing the efficacy of an academic hearing peer tutor for a profoundly deaf student. *American Annals of the Deaf, 139*, 415–419.

Burnett, L. McLaughlin, T. F., & Hunsaker, D. (1978). Instruction-following behavior of an entire third-grade classroom: Effects of a timing device, public posting and teacher praise. *Behavioral Engineering, 5*, 37–40.

*Cameron, M. J., Shapiro, R. L., & Ainsleigh, S. A. (2005). Bicycle riding: Pedaling made possible through positive behavioral interventions. *Journal of Positive Behavior Interventions, 7*, 153–159.

*Cannella-Malone, H. I., DeBar, R. M., & Sigafoos, J. (2009). An examination of preference for augmentative and alternative communication devices with two boys with significant intellectual disabilities. *Augmentative Alternative Communication, 25*, 262–273.

*Carson, L., Moosa, T., Theurer, J., & Cardy, J. O. (2012). The collateral effects of PECS training on speech development in children with autism. *Canadian Journal of Speech-Language Pathology and Audiology, 36*, 182–195.

*Cavalari, R. N. S., DuBard, M., Luiselli, J. K., & Birtwell, K. (2013). Teaching an adolescent with autism and intellectual disability to tolerate routine medical examination: effects of a behavioral compliance training package. *Clinical Practice in Pediatric Psychology, 1*, 121–128.

*Creech, J., & Golden, J. A. (2009). Increasing Braille practice and reading comprehension in a student with visual impairment and moderate mental retardation: An initial study and follow-up. *Journal of Developmental and Physical Disabilities, 21*, 225–233.

*Davis, D. H., Gagne, P., Fredrick, L. D., Alberto, P. A., Waugh, R. E., & Haardorfer, R. (2012). Augmenting visual analysis in single-case research with hierarchical linear modeling. *Behavior Modification, 37*, 62–89.

*De Luca, R. V., & Holborn, S. W. (1992). Effects of a variable-ratio reinforcement schedule with changing criteria on exercise in obese and nonobese boys. *Journal of Applied Behavior Analysis, 25*, 671–679.

*Dillenburger, K., & Keenan, M. (2003). Using hypnosis to facilitate direct observation of multiple tics and self-monitoring in a typically developing teenager. *Behavior Therapy, 34*, 117–125.

*Donohue, M. M., Casey, L. B., Bicard, D. F., & Bicard, S. E. (2012). Effects of differential reinforcement of short latencies on responce latency, task completion, and accuracy of an adolescent with Autism. *Education and Training in Autism and Developmental Disabilities, 47*, 97–108.

*Easterbrooks, S. R., & Stoner, M. (2006). Using a visual tool to increase adjectives in the written language of students who are deaf or hard of hearing. *Communication Disorders Quarterly, 27*, 95–123.

*Edinger, J. D., Nelson, W. M., III, Davidson, K. M., & Wallace, J. (1978). Modification of smoking behaviors in a correctional institution. *Journal of Clinical Psychology, 34*, 991–998.

*Eidson, C. E., Jr. (1989). The effect of behavioral music therapy on the generalization of interpersonal skills from sessions to the classroom by emotionally handicapped middle school students. *Journal of Music Therapy, 26*, 206–221.

*Fienup, D. M., & Doepke, K. (2008). Evaluation of a changing criterion intervention to increase fluent responding with an elementary age student with autism. *International Journal of Behavioral Consultation and Therapy, 4*, 297–303.

*Fienup, D. M., Mudgal, D., & Pace, G. (2013). Increasing money-counting skills with a student with brain injury: skill and performance deficits. *Brain Injury, 27,* 366–376.

*Fitterling, J. M., Martin, J. E., Gramling, S., Cole, P., & Milan, M. A. (1988). Behavioral management of exercise training in vascular headache patients: An investigation of exercise adherence and headache activity. *Journal of Applied Behavior Analysis, 21,* 9–19.

*Flood, W. A., & Wilder, D. A. (2004). The use of differential reinforcement and fading to increase time away from a caregiver in a child with separation anxiety disorder. *Education and Treatment of Children, 27,* 1–8.

*Fordyce, W. E., Shelton, J. L., & Dundore, D. E. (1982). The modification of avoidance learning pain behaviors. *Journal of Behavioral Medicine, 5*, 405–415.

*Foxx, R. M., & Rubinoff, A. (1979). Behavioral treatment of caffeinism: Reducing excessive coffee drinking. *Journal of Applied Behavior Analysis, 12*, 335–344.

*Fredrick, L. D., Davis, D. H., Alberto, P. A., & Waugh, R. E. (2013). From initial phonics to functional phonics: Teaching word-analysis skills to students with moderate intellectual disability. *Education and Training in Autism and Developmental Disabiliites, 48,* 49–66.

*Ganz, J. B., & Flores, M. M. (2008). Effects of the use of visual strategies in play groups for children with autism spectrum disorders and their peers. *Journal of Autism and Developmental Disorders, 38*, 926–940.

*Ganz, J. B., & Flores, M. M. (2009). The effectiveness of direct instruction for teaching language to children with autism spectrum disorders: Identifying materials. *Journal of Autism and Developmental Disabilities, 39*, 75–83.

*Ganz, J. B., & Sigafoos, J. (2005). Self-monitoring: Are young adults with MR and autism able to utilize cognitive strategies independently? *Education and Training in Developmental Disabilities, 40*, 24–33.

*Gardner, W. I., Clees, T. J., & Cole, C. L. (1983). Self-management of disruptive verbal ruminations by a mentally retarded adult. *Applied Research in Mental Retardation, 4*, 41–58.

*Gorski, J. A., & Westbrook, A. C. (2002). Use of differential reinforcement to treat medical non-compliance in a paediatric patient with leukocyte adhesion deficiency. *Pediatric Rehabilitation, 5*, 29–35.

*Grey, I., Healy, O., Leader, G., & Hayes, D. (2009). Using a Time Timer to increase appropriate waiting behavior in a child with developmental disabilities. *Research in Developmental Disabilities, 30*, 359–366.

*Hagopian, L. P., & Slifer, K. J. (1993). Treatment of separation anxiety disorder with graduated exposure and reinforcement targeting school attendance: A controlled case study. *Journal of Anxiety Disorders*, *7*, 271–280.

*Hamilton, B., & Snell, M. E. (1993).Using the milieu approach to increase spontaneous communication book use across environments by an adolescent with autism. *Augmentative and Alternative Communication, 9*, 259–272.

*Hartmann, D. P., & Hall, R. V. (1976). The changing criterion design. *Journal of Applied Behavior Analysis*, *9*, 527–532.

*Hastie, P. A., & Sharpe, T. (2006). Introducing a changing-criterion design to hold students accountable in structured physical activity settings. *Journal of Evidence-Based Practices for Schools*, *7*, 73–88.

*Himadi, B., Osteen, F., Kaiser, A. J., & Daniel, K. (1991). Assessment of delusional beliefs during the modification of delusional verbalizations. *Behavioral Residential Treatment, 6*, 355–366.

*Hinerman, P. S., Jenson, W. R., Walker, G. R., & Petersen, P. B. (1982). Positive practice overcorrection combined with additional procedures to teach signed words to an Autistic child. *Journal of Autism and Developmental Disorders, 12*, 253–263.

*Johnston, R. J., & McLaughlin, T. F. (1982). The effects of free time on assignment completion and accuracy in arithmetic: A case study. *Education and Treatment of Children, 5*, 33–40.

*Kahng, S., Boscoe, J. H., & Shannon, B. (2003). The use of escape contingency and a token economy to increase food acceptance. *Journal of Applied Behavior Analysis, 36*, 349–353.

*Kern, L., Starosta, K. M., Cook, C. R., Bambara, L. M., & Gresham, F. R. (2007). Functional assessment-based intervention for selective mutism. *Behavioral Disorders, 32*, 94–108.

*Koegel, R. L., Openden, D., & Koegel, L. K. (2004). A systematic desensitization paradigm to treat hypersensitivity to auditory stimuli in children with autism in family contexts. *Research & Practice for Persons with Severe Disabilities, 29*, 122–134.

*Kowalewicz, E. A., & Coffee, G. (2014). Mystery Motivator: A Tier 1 classroom behavioral intervention. *School Psychology Quarterly, 29*, 138–156.

*Lamm, N., & Greer, R. D. (1988). Induction and maintenance of swallowing responses in infants with dysphagia. *Journal of Applied Behavior Analysis, 21*, 143–156.

*Lane, K. L., Capizzi, A, M., Fisher, M. H., & Ennis, R. P. (2012). Secondary prevention efforts at the middle school level: an application of the behavior education program. *Education and Treatment of Children*, *35*, 51–90.

*Lane, K. L., Rogers, L. A., Parks, R. J., Weisenbach, J. L., Mau, A. C., Merwin, M. T., & Bergman, W. A. (2007). Function-based interventions for students who are nonresponsive to primary and secondary prevention efforts: Illustrations at the elementary and middle school levels. *Journal of Emotional and Behavioral Disorders, 15*, 169–183.

*Layton, C. A., & Koenig, A. J. (1998). Increasing reading fluency in elementary students with low vision through repeated readings. *Journal of Visual Impairment & Blindness*, *92*, 276–292.

*Lee, C. Y. Q., Anderson, A., & Moore, D. W. (2014). Using video modeling to toilet train a child with Autism. *Journal of Developmental and Physical Disability, 26*, 123–134.

*Linscheid, T. R., Tarnowski, K. J., Rasnake, L. K., & Brams, J. S. (1987). Behavior treatment of food refusal in a child with short-gut syndrome. *Journal of Pediatric Psychology*, *12*, 451–459.

*Luiselli, J. K. (2000). Cueing, demand fading, and positive reinforcement to establish self-feeding and oral consumption in a child with chronic food refusal. *Behavior Modification, 24*, 348–358.

*Lutzker, J. R., Campbell, R. V., & Watson-Perczel, M. (1984). Using the case study method to treat several problems in a family indicated for child neglect. *Education and Treatment of Children*, *7*, 315–333.

*Marcus, A., Sinnott, B., Bradley, S., & Grey, I. (2010). Treatment of idiopathic toe-walking in children with Autism using GaitSpot Auditory Speakers and simplified habit reversal. *Research in Autism Spectrum Disorders, 4*, 260–267.

*Martella, R. C., Leonard, I. J., Marchand-Martella, N. E., & Agran, M. (1993). Self-monitoring negative statements. *Journal of Behavioral Education, 31*, 77–86.

*Martens, B. K., Eckert, T. L., Begeny, J. C., Lewandowski, L. J., DiGennaro, F. D., Montarello, S. A., Arbolino, L. A., Reed, D. D., & Fiese, B. H. (2007). Effects of a fluency-building program on the reading performance of low-achieving second and third grade students. *Journal of Behavioral Education, 16,* 38–53.

*McCartney, E. J., Anderson, C. M., & English, C. L. (2005). Effect of brief clinic-based training on the ability of caregivers to implement escape extinction. *Journal of Positive Behavior Support, 7,* 18–33.

*Meyers, A. W., Artz, L. M., & Craighead, W. E. (1976). The effects of instructions, incentive, and feedback on a community problem: Dormitory noise. *Journal of Applied Behavior Analysis, 9*, 445–457.

*Mizes, J. S. (1985). The use of contingent reinforcement in the treatment of a conversion disorder: A multiple baseline study. *Journal of Behavior Therapy and Experimental Psychiatry, 16*, 341–345.

*Morgan, R. L., & Hiebert, R. (1982). The changing-criterion design: Assessment of running behavior of a handicapped youth. *Psychological Reports, 50*, 1287–1993.

*Mruzek, D. W., Cohen, C., & Smith, T. (2007). Contingency contracting with students with Autism Spectrum Disorders in a public school setting. *Journal of Developmental and Physical Disabilities, 19*, 103–114.

*Nes Ferrara, S. L. (2005). Reading fluency and self-efficacy: A case study. *International Journal of Disability, Development and Education, 52*, 215–231.

*Nes, S. L. (2003). Using paired reading to enhance the fluency skills of less-skilled readers. *Reading Improvement, 40*, 179–192.

*Nicol, N., & Hantula, D. A. (2001). Decreasing delivery drivers' departure times. *Journal of Organizational Behavior Management, 21*, 105–116.

*Noles, K. O., Epstein, L. H., & Jones, H. T. (1976). Instruction control of sleep reduction. *Perceptual and Motor Skills, 42*, 669–670.

*O'Connor, A. S., Prieto, J., Hoffman, B., DeQuinzio, J. A., & Taylor, B. A. (2011). A stimulus control procedure to decrease motor and vocal stereotypy. *Behavioral Interventions, 26*, 231–242.

*Ohtake, Y., Wehmeyer, M., Uchida, N., Nakaya, A., & Yanagihara, M. (2010). Enabling a prelinguistic communicator with Autism to use picture card as a strategy for repairing listener misunderstandings: A case study. *Education and Training in Autism and Developmental Disabilities, 45*, 410–421.

*Osborne, M. L., & Himadi, B. (1990). Evaluation of a shaping procedure with the changing-criterion design. *Behavioral Residential Treatment, 5*, 75–81.

*Pailthorpe, W. K., & Ralph, A. (1998). Time-out as a means of shaping whole-task completion as a precursor to establishing rule-following behaviour with a severely noncompliant preschool child. *Behaviour Change, 15*, 50–61.

*Paniagua, F. A. (1992). Verbal-nonverbal correspondence training with ADHD children. *Behavior Modification, 16*, 226–252.

*Paniagua, F. A., & Saeed, M. A. (1987). Labeling and functional language in a case of psychological mutism. *Journal of Behavior Therapy and Experimental Psychiatry, 18*, 259–267.

*Pattillo, S. T., Heller, K. W., & Smith, M. (2004). The impact of a modified repeated-reading strategy paired with optical character recognition on the reading rates of students with visual impairments. *Journal of Visual Impairment & Blindness, 98*, 28–46.

*Pavchinski, P. Evans, J. H., & Bostow, D. E. (1989). Increasing word recognition and math ability in a severely learning-disabled student with token reinforcers. *Psychology in the Schools, 26*, 397–411.

*Ricciardi, J. N., Luiselli, J. K., & Camare, M. (2006). Shaping approach responses as intervention for specific phobia in a child with Autism. *Journal of Applied Behavior Analysis, 39*, 445–448.

*Rush, D. B., & Ayllon, T. (1984). Peer behavioral coaching: Soccer. *Journal of Sport Psychology, 6*, 325–334.

*Sanok, R. L., & Ascione, F. R. (1978). The effects of reduced time limits on prolonged eating behavior. *Journal of Behavior Therapy and Experimental Psychiatry, 9*, 177–179.

*Schleien, S. J., Wehman, P., & Kiernan, J. (1981). Teaching leisure skill to severely handicapped adults: An age-appropriate darts game. *Journal of Applied Behavior Analysis, 14*, 513–519.

*Schloss, P. J., Sedlak, R. A., Elliot, C., & Smothers, M. (1982). Application of the changing criterion design in special education. *The Journal of Special Education, 16*, 359–367.

*Schumacher, B. I., & Rapp, J. T. (2011). Increasing compliance with haircuts in a child with autism. *Behavioral Interventions, 26*, 67–75.

*Scott, D., Scott, L. M., & Howe, B. L. (1998). Training anticipation for intermediate tennis players. *Behavior Modification, 22*, 243–261.

*Shaw, R., & Simms, T. (2009). Reducing attention-maintained behavior through the use of positive punishment, differential reinforcement of low rates, and response marking. *Behavioral Interventions, 24*, 249–263.

*Shrestha, A., Anderson, A., & Moore, D. W. (2013). Using point-of-view video modeling and forward chaining to teach a functional self-help skill to a child with Autism. *Journal of Behavioral Education, 22*, 157–167.

*Singh, N. N., Lancioni, G. E., Singh, A. N. A., Winton, A. S. W., Singh, A. D. A., & Singh, J. (2011). A mindfulness-based health wellness program for individuals with Prader-Willi syndrome. *Journal of Mental Health Research in Intellectual Disabilities, 4*, 90–106.

*Skinner, C. H., Skinner, A. L., & Armstrong, K. J. (2000). Analysis of a client-staff-developed shaping program designed to enhance reading persistence in an adult diagnosed with schizophrenia. *Psychiatric Rehabilitation Journal, 24*, 52–58.

Smith, M. A., Schloss, P. J., & Israelite, N. I. (1986). Evaluation of a simile recognition treatment program for hearing-impaired students. *Journal of Speech and Hearing Disorders, 51*, 134–139.

*Sulzer-Azaroff, B., Loafman, B., Merante, R. J., & Hlavacek, A. C. (1990). Improving occupational safety in a large industrial plant: A systematic replication. *Journal of Organizational Behavior Management, 11*, 99–121.

*Susa, C., & Schlinger, H. D. (2012). Using a lag schedule to increase variability of verbal responding in an individual with Autism. *The Analysis of Verbal Behavior, 28*, 125–130.

*Taggart, A. C., Taggart, J., & Siedentop, D. (1986). Effects of a home-based activity program: A study with low fitness elementary school children. *Behavior Modification, 10*, 487–507.

*Tereshko, L., MacDonald, R., & Ahearn, W. H. (2010). Strategies for teaching children with Autism to imitate response chains using video modeling. *Research in Autism Spectrum Disorder, 4*, 479–489.

*Todd, T., Reid, G., & Butler-Kisber, L. (2010). Cycling for students with ASD: Self-regulation promotes sustained physical activity. *Adapted Physical Activity Quarterly, 27*, 226–241.

*Wallace, C. J., & Davis, J. R. (1974). Effects of information and reinforcement on the conversational behavior of chronic psychiatric patient dyads. *Journal of Consulting and Clinical Psychology, 42*, 656–662.

*Warnes, E., & Allen, K. D. (2005). Biofeedback treatment of paradoxical vocal fold motion and respiratory distress in an adolescent girl. *Journal of Applied Behavior Analysis, 38*, 529–532.

*Watson, T. S., Allen, S. J., & Allen, K. D. (1993). Ventricular fold dysphonia: Application of biofeedback technology to a rare voice disorder. *Behavior Therapy, 24*, 439–446.

*Weissman, W., Glasgow, R., Biglan, A., & Lichtenstein, E. (1987). Developmental and preliminary evaluation of a cessation program for adolescent smokers. *Psychology of Addictive Behaviors, 1*, 84–91.

*Williamson, D. A., Lawson, O. D., Bennett, S. M., & Hinz, L. (1989). Behavioral treatment of night bingeing and rumination in an adult case of bulimia nervosa. *Journal of Behavior Therapy and Experimental Psychiatry*, *20*, 73–77.

*Wolfe, L. H., Heron, T. E., & Goddard, Y. L. (2000). Effects of self-monitoring on the on-task behavior and written language performances of elementary students with learning disabilities. *Journal of Behavioral Education, 10*, 49–73.

*Wolff, J. J., & Symons, F. J. (2012). An evaluation of multi-component exposure treatment of needle phobia in an adult with Autism and Intellectual Ability. *Journal of Applied Research in Intellectual Disabilities, 26*, 344–348.

Appendix B: References from Our Own Bibliographic Search

Brogan, K. M., Falligant, J. M., & Rapp, J. T. (2017). Interdependent group contingencies decrease adolescents' disruptive behaviors during group therapy: A practitioner's demonstration. *Behavior Modification*, *41*, 405–421.

Cervantes, C. M., & Porretta, D. L. (2013). Impact of after school programming on physical activity among adolescents with visual impairments. *Adapted Physical Activity Quarterly*, *30*, 127–146.

Cook, J. L., Rapp, J. T., & Schulze, K. A. (2015). Differential negative reinforcement of other behavior to increase wearing of a medical bracelet. *Journal of Applied Behavior Analysis, 48*, 901–906.

Cromartie, R. S., Flood, W. A., & Luiselli, J. K. (2014). Graduated exposure and compliance training intervention for blood draw avoidance and refusal in a woman with intellectual disability and schizoaffective disorder. *Journal of Mental Health Research in Intellectual Disabilities*, *7*, 95–103.

Facon, B., Sahiri, S., & Rivière, V. (2008). A controlled single-case treatment of severe long-term selective mutism in a child with mental retardation. *Behavior Therapy*, *39*, 313–321.

Fischer, A. J., Luiselli, J. K., & Dove, M. B. (2015). Effects of clinic and in-home treatment on consumption and feeding-associated anxiety in an adolescent with avoidant/restrictive food intake disorder. *Clinical Practice in Pediatric Psychology*, *3*, 154–166.

Gann, C. J., & Kunnavatana, S. S. (2016). A preliminary study in applying the function-based intervention decision model in consultation to increase treatment integrity. *Education and Treatment of Children*, *39*, 445–466.

Gentry, J. A., & Luiselli, J. K. (2008). Treating a child's selective eating through parent implemented feeding intervention in the home setting. *Journal of Developmental and Physical Disabilities*, *20*, 63–70.

Guck, T. P., Burke, R. V., Rainville, C., Hill-Taylor, D., & Wallace, D. P. (2015). A brief primary care intervention to reduce fear of movement in chronic low back pain patients. *Translational Behavioral Medicine, 5,* 113–121.

Hodges, A., Davis, T., Crandall, M., Phipps, L., & Weston, R. (2017, May 23). Using shaping to increase foods consumed by children with autism. *Journal of Autism and Developmental Disorders*. Advance online publication. http://dx.doi.org.sire.ub.edu/10.1007/s10803-017-3160-y

Ingham, R. J., Sato, W., Finn, P., & Belknap, H. (2001). The modification of speech naturalness during rhythmic stimulation treatment of stuttering. *Journal of Speech, Language, and Hearing Research*, *44*, 841–852.

Kadey, H. J., Roane, H. S., Diaz, J. C., & Merrow, J. M. (2013). An evaluation of chewing and swallowing for a child diagnosed with autism. *Journal of Developmental and Physical Disabilities, 25*, 343–354.

Kreibich, S. R., Chen, M., & Reichle, J. (2015). Teaching a child with autism to request breaks while concurrently increasing task engagement. *Language, Speech, and Hearing Services in Schools*, *46*, 256–265.

Laprime, A. P., & Dittrich, G. A. (2014). An evaluation of a treatment package consisting of discrimination training and differential reinforcement with response cost and a social story on vocal stereotypy for a preschooler with autism in a preschool classroom. *Education and Treatment of Children*, *37*, 407–430.

Luiselli, J. K., Duncan, N. G., Keary, P., Nelson, E. G., Parenteau, R. E., & Woods, K. E. (2013). Behavioral coaching of track athletes with developmental disabilities: Evaluation of sprint performance during training and Special Olympics competition. *Journal of Clinical Sport Psychology, 7*, 264–274.

McDaniel, S. C., & Bruhn, A. L. (2016). Using a changing-criterion design to evaluate the effects of check-in/check-out with goal modification. *Journal of Positive Behavior Interventions, 18*, 197–208.

Neufeld, V., Law, K. C., & Lucyshyn, J. M. (2014). Integrating best practices in positive behavior support and clinical psychology for a child with autism and anxiety-related problem behavior: A clinical case study. *Canadian Journal of School Psychology, 29*, 258–276.

O'Handley, R. D., Dadakhodjaeva, K., Radley, K. C., & Dart, E. H. (2016). Promoting independent ambulation: A case study of an elementary school student with developmental disabilities. *Research in Developmental Disabilities, 56*, 153–159.

Orosco, M. J. (2014). A math intervention for third grade Latino English language learners at risk for math disabilities. *Exceptionality, 22*, 205–225.

Parnell, A. M., Lorah, E. R., Karnes, A., & Schaefer-Whitby, P. (2017). Effectiveness of job aids and post performance review on staff implementation of discrete trial instruction. *Journal of Organizational Behavior Management, 37*, 207–220.

Pérez González, L. A., & Williams, G. (2006). Comprehensive program for teaching skills to children with autism. *Psychology in Spain, 10*, 37–51.

Russel, C. S., Allday, R. A., & Duhon, G. J. (2015). Effects of increasing distance of a one-on-one paraprofessional on student engagement. *Education and Treatment of Children, 38*, 193–210.

Savaiano, M. E., & Hatton, D. D. (2013). Using repeated reading to improve reading speed and comprehension in students with visual impairments. *Journal of Visual Impairment & Blindness, 107*, 93.

Stasolla, F., Perilli, V., Boccasini, A., Caffò, A. O., Damiani, R., & Albano, V. (2016). Enhancing academic performance of three boys with Autism Spectrum Disorders and intellectual disabilities through a computer-based program. *Life Span and Disability, 19*, 153–183.

Voulgarakis, H., & Forte, S. (2015). Escape extinction and negative reinforcement in the treatment of pediatric feeding disorders: A single case analysis. *Behavior Analysis in Practice, 8*, 212–214.

Weinkauf, S. M., Zeug, N. M., Anderson, C. T., & Ala'i-Rosales, S. (2011). Evaluating the effectiveness of a comprehensive staff training package for behavioral interventions for children with autism. *Research in Autism Spectrum Disorders, 5*, 864–871.

List of Excluded Studies

For instance, from the references in Cromartie, Flood, and Luiselli (2014), an article by Grider, Luiselli, and Turcotte-Shamski (2012) was identified and it used a design resembling a CCD, but did not name it accordingly and thus Grider et al. (2012) was not included in the current review. The same is the case for the study by Fairbanks, Sugai, Guardino, and Lathrop (2007) identified from the references in McDaniel and Bruhn (2006), and for the study by Manella and Varni (1984) identified from the references in O'Handley, Dadakhodjaeva, Radley, and Dart (2016).

Fairbanks, S., Sugai, G., Guardino, D., & Lathrop, M. (2007). Response to intervention: Examining classroom behavior support in second grade. *Exceptional Children, 73*, 288–310.

Grider, B., Luiselli, J. K., & Turcotte-Shamski, W. (2012). Graduated exposure, positive reinforcement, and stimulus distraction in a compliance-with-blood-draw intervention for an adult with autism. *Clinical Case Studies, 11*, 253–260.

Manella, K. J., & Varni, J. W. (1984). Behavioral treatment of ambulatory function in a child with myelomeningocele: A case report. *Physical Therapy, 64*, 1536–1539.

Excluded due to using non-human participants:

Manabe, K., Dooling, R. J., & Takaku, S. (2013). Differential reinforcement of an approach response in zebrafish (Danio rerio). *Behavioural Processes, 98*, 106–111.

Slater, C., & Dymond, S. (2011). Using differential reinforcement to improve equine welfare: Shaping appropriate truck loading and feet handling. *Behavioural Processes*, *86*, 329–339.

Appendix C: Figures representing data from the literature review

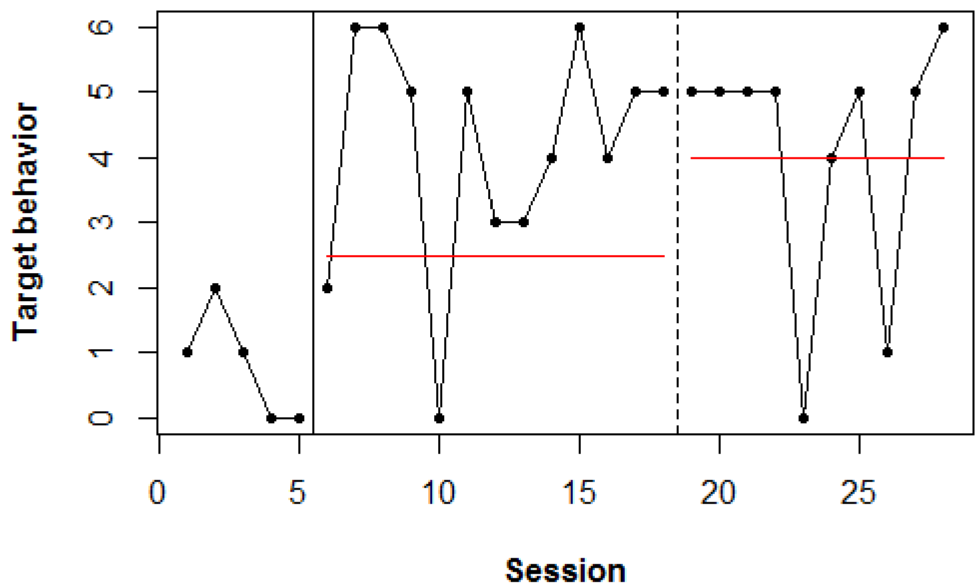


Figure A1. Data gathered by S. J. Allen and Kramer (1990) on the number of target behaviors observed daily, during baseline (before the solid vertical line) and under two criterion levels during the intervention phase. The solid horizontal lines in the intervention sub-phases represent the criterion levels.

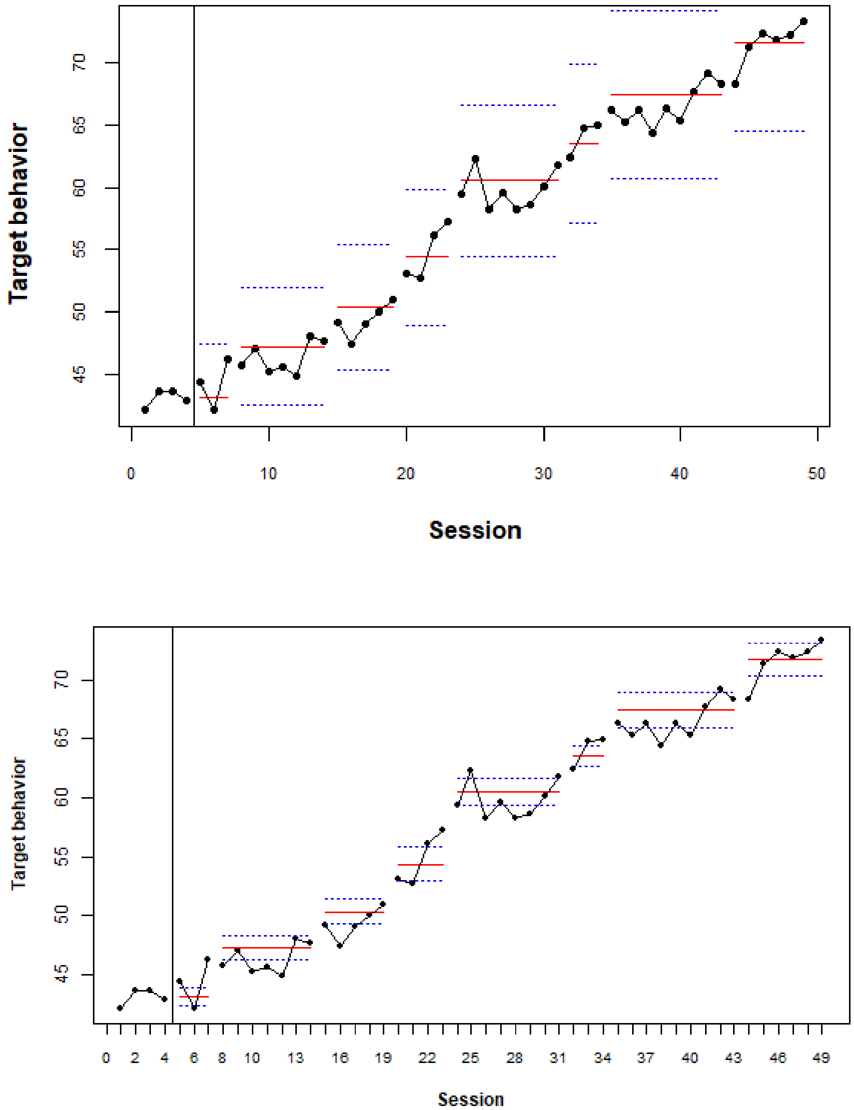


Figure A2. Data gathered by Facon, Sahiri and Revière (2008) on speech loudness measured in decibels, during baseline (before the solid vertical line) and under eight criterion levels during the intervention phase. The Y-axis is the original one used by Facon et al. (2008). The solid horizontal lines in the intervention sub-phases represent the criterion levels. The dotted horizontal lines represent acceptable ranges: upper panel - based on the criterion plus/minus 10%; lower panel – based on our proposal using values of p = 0.5, 0.3, 0.5, 0.5, 0.25, 0.5, 0.5, and 0.5 (chosen by trial and error to produce nonoverlapping ranges similar in absolute spread).

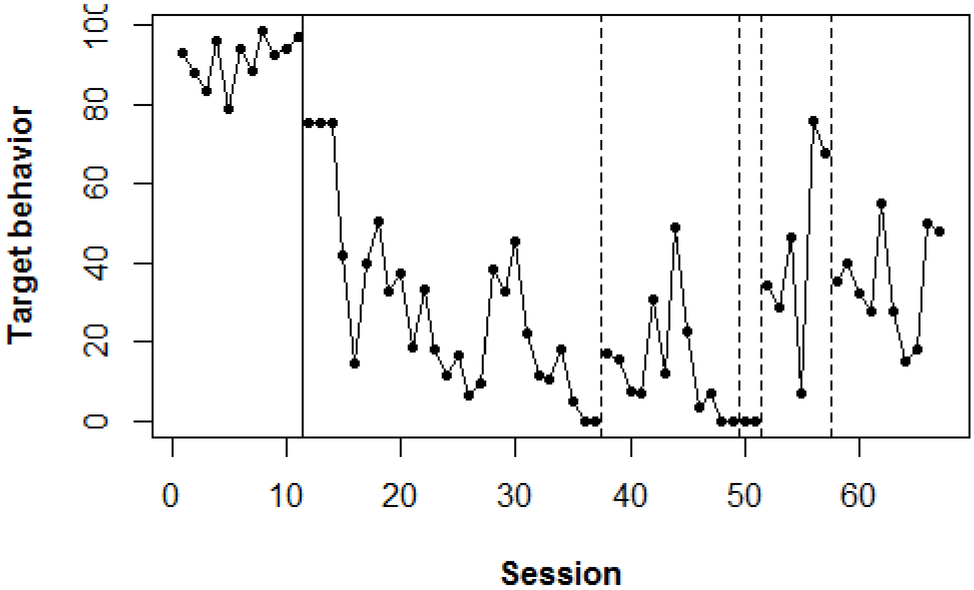


Figure A3. Data gathered by Marcus, Sinnott Bradley, and Grey (2010) on the percentage of toe-walking across ten minutes per day by a participant called Damien, during baseline (before the solid vertical line) and under five criterion levels during the intervention phase. The mastery criterion was two consecutive days with zero target behavior.

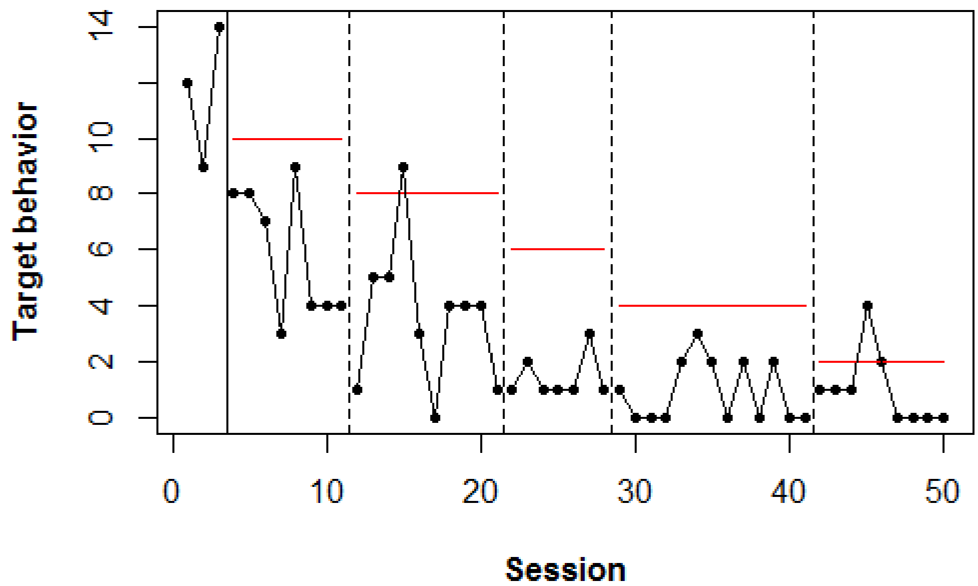


Figure A4. Data gathered by Shaw and Simms (2009) on the frequency of disruptive episodes by a participant called Andrew, during baseline (before the solid vertical line) and under five criterion levels during the intervention phase. The solid horizontal lines in the intervention sub-phases represent the criterion levels.

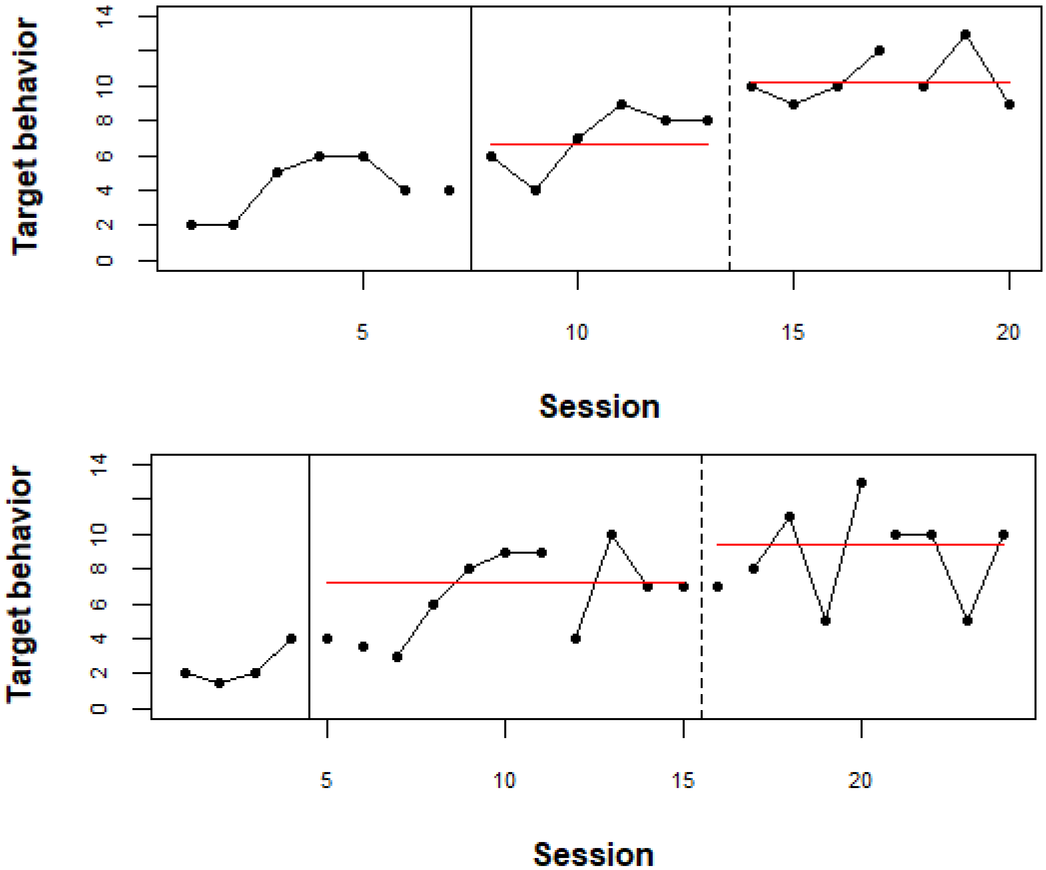


Figure A5. Data gathered by Todd, Reid, and Butler-Kisber (2010) on distance cycled, in number of circuits completed per session, by participants called Lisa (upper panel) and Daniel (lower panel), during baseline (before the solid vertical line) and under two criterion levels during the intervention phase. The solid horizontal lines in the intervention sub-phases represent the criterion levels. The discontinuities in the within-phase lines are represent missed sessions in between.

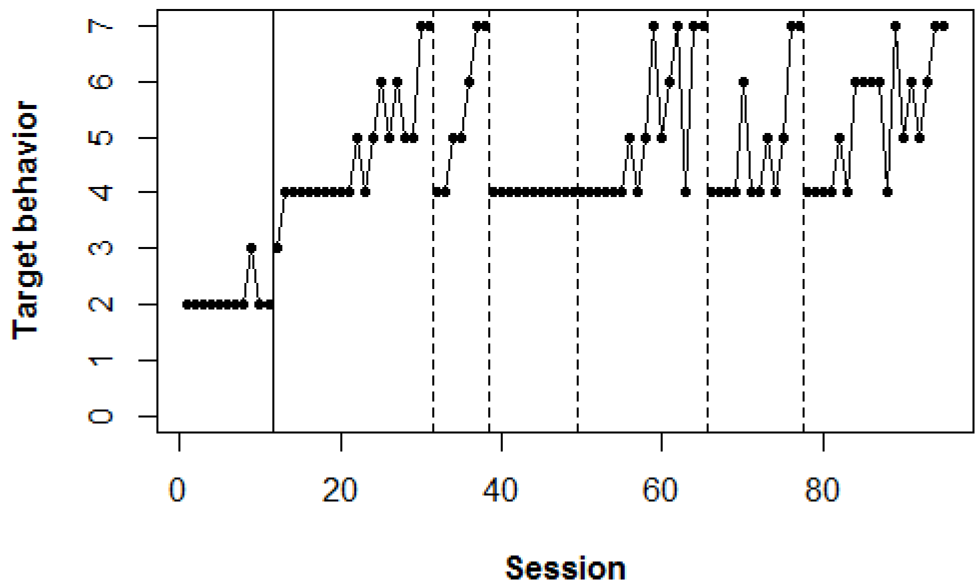


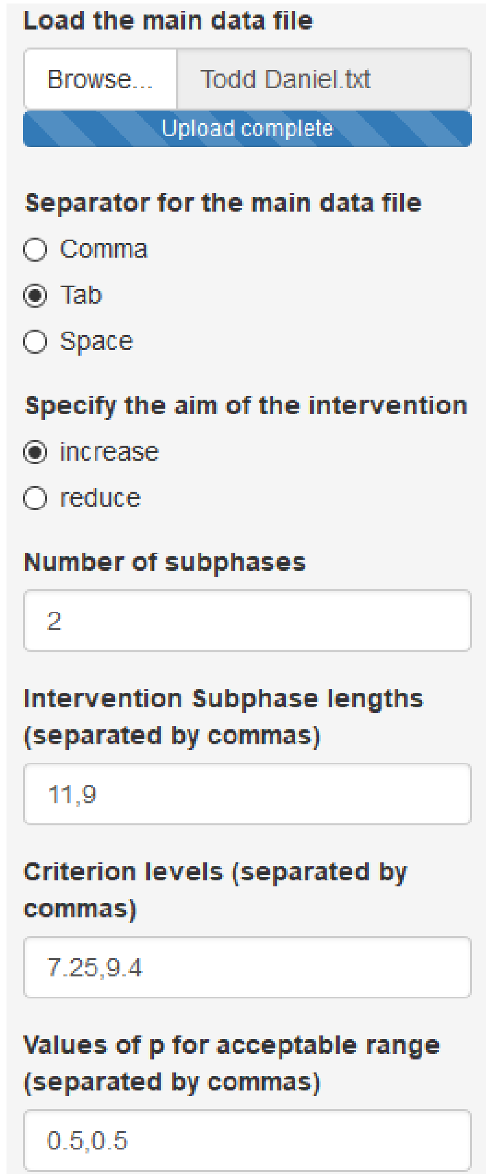
Figure A6. Data gathered by Schleien, Wehman, and Kiernan (1981) on the number of steps performed independently by a participant called Roselyn, during baseline (before the solid vertical line) and under six criterion levels during the intervention phase. The mastery criterion was two consecutive completions of the seven steps.

Appendix D: Web-Based Application

1. Organize the data files, as shown in the example below.

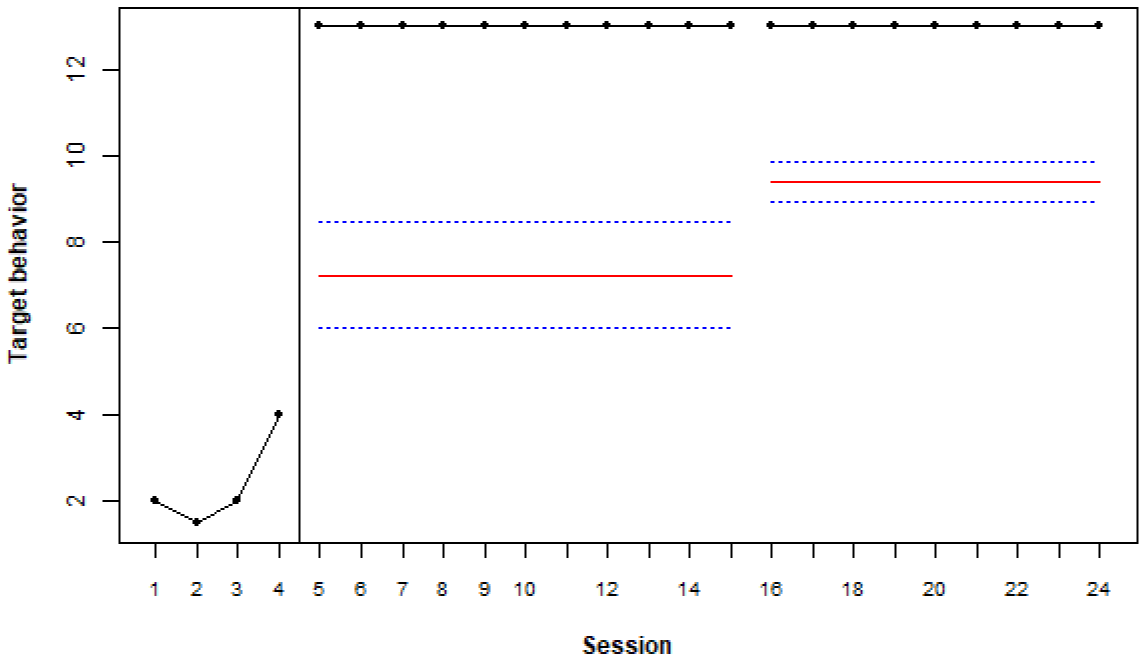
Empty data file: use to define the acceptable ranges after gathering the baseline data, but before gathering the intervention phase data. The data for the intervention sub-phase are made up and set to some relatively large value		Actual data file: use to obtain the result	
score	phase	score	phase
2.00	A	2.00	A
1.50	A	1.50	A
2.00	A	2.00	A
4.00	A	4.00	A
13.00	B1	4.00	B1
13.00	B1	3.50	B1
13.00	B1	3.00	B1
13.00	B1	6.00	B1
13.00	B1	8.00	B1
13.00	B1	9.00	B1
13.00	B1	9.00	B1
13.00	B1	4.00	B1
13.00	B1	10.00	B1
13.00	B1	7.00	B1
13.00	B1	7.00	B1
13.00	B2	7.00	B2
13.00	B2	8.00	B2
13.00	B2	11.00	B2
13.00	B2	5.00	B2
13.00	B2	13.00	B2

2. Enter the relevant data, as shown in the screenshots of the Shinny web-based application from https://manolov.shinyapps.io/ChangingCriterion.

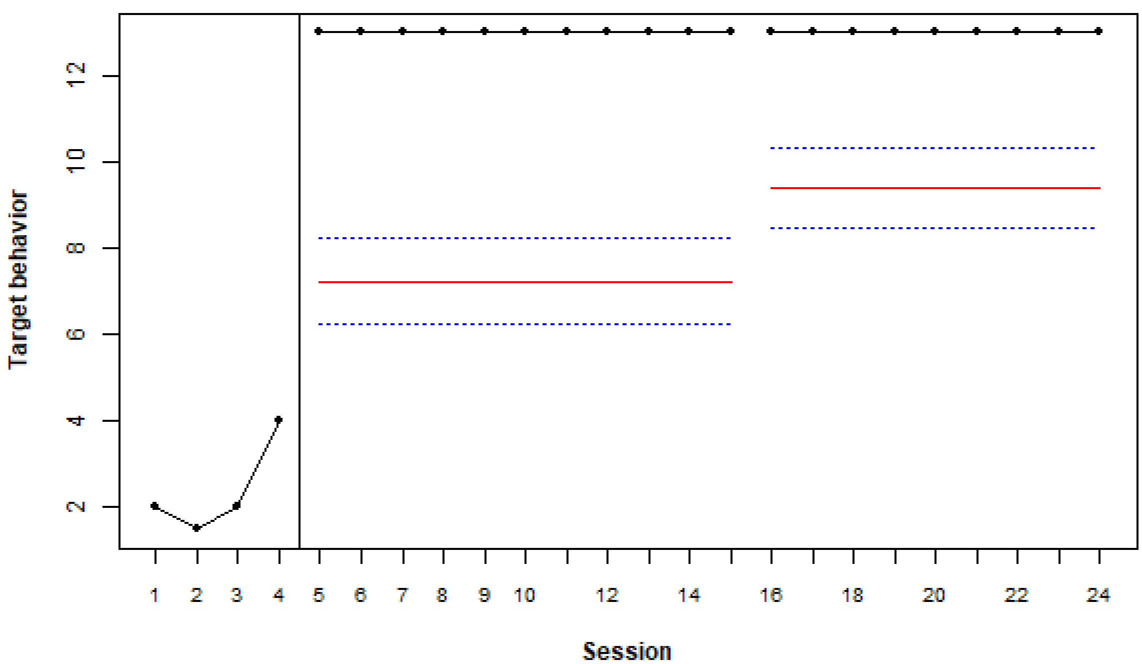


The following information is entered representing the participant Daniel from Todd et al. (2010): (a) number of subphases: 2; (b) intervention subphase lengths: 11,9; (c) criterion levels: 7.25,9.4; (d) example initial values of p: 0.5,0.5.

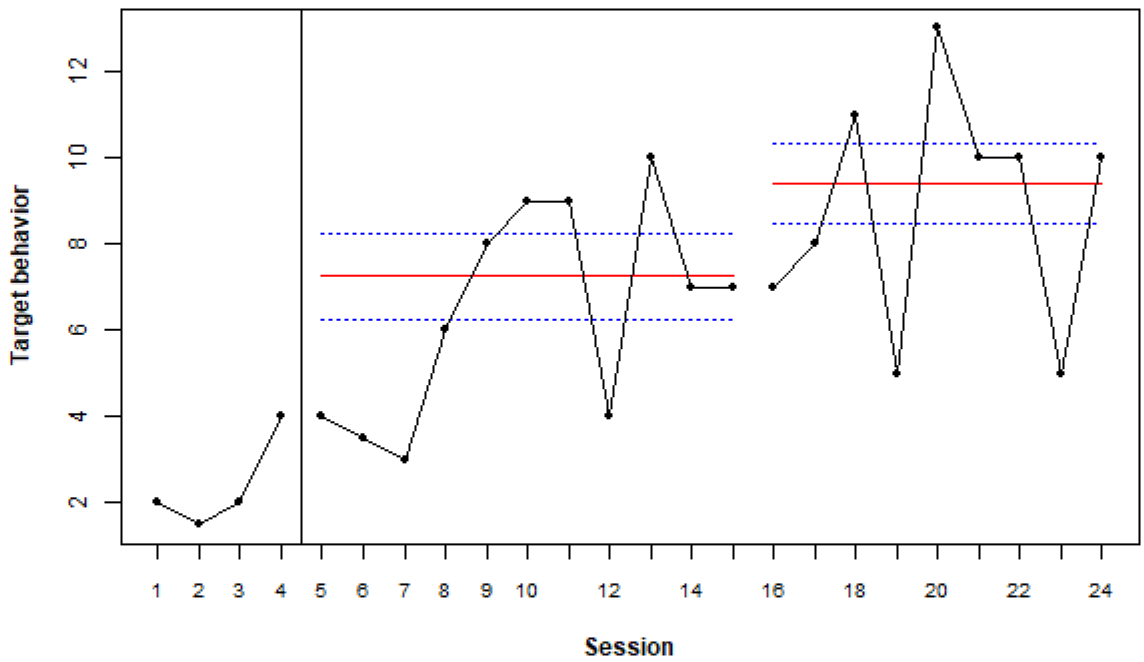
3. Adjust the values of p, by trial and error to obtain acceptable ranges, which are reasonable, according to the criterion of the researcher. Using the initial values:0.5 and 0.5, the following ranges are obtained:



For the current example, we chose p: 0.4 and 0.8



4. Obtain the results with the actual data: Graphical representation



5. Obtain the results with the actual data: Numerical summary

```
Nonoverlap of all pairs
Comparing B1 and  B2 . NAP= 0.76

ACCEPTABLE RANGE: Percentage conforming data
Frequency conforming data per criterion level:
For criterion level 1 : 3 values
For criterion level 2 : 3 values

Percentage conforming data per criterion level:
For criterion level 1 : 27.27 %
For criterion level 2 : 33.33 %
```