In search of scope: A response to Ruiz et al. (2020)

Hulbert-Williams, L., Pendrous, R., Hochard, K. D., & Hulbert-Williams, N. J.

Centre for Contextual Behavioural Science, School of Psychology, University of Chester, UK

Author Note

This is an invited response to:

Ruiz, F. J., Luciano, C., & Sierra, M. A. (2020). A systematic and critical response to Pendrous et al. (2020) replication study. *Journal of Contextual Behavioral Science*, *17*, 39-45. https://doi.org/10.1016/j.jcbs.2020.04.011

Correspondence concerning this article should be addressed to Rosina Pendrous, ropendrous@gmail.com; School of Psychology, University of Chester, Parkgate Road, Chester, CH1 4BJ.

Highlights

- Replicability is a core feature of the scientific process.
- Both replications and responses to replications should be carefully communicated.
- Ruiz et al. critiqued Pendrous et al. for methodological differences.
- These differences are newly implied to be moderators of the effect.
- If Ruiz et al. are correct, theoretical scope may become limited.

Abstract

Deliberate and explicit replication attempts are becoming more common across the behavioral sciences. Whilst replicability has been recognized as a core feature of science for decades (if not centuries), the directness of today's replication work requires us to consider carefully how we communicate our research and how we conceptualize our theories in light of differing findings. This paper uses a concrete example to make a number of suggestions for how we, as a scientific community, ought to engage with replication attempts. Within Relational Frame Theory (RFT) there is a growing body of applied research on the effective use of metaphors to increase tolerance of aversive states. We conducted a replication of an earlier experimental analogue study (2020, this journal) and failed to find the specified effect. Ruiz et al. (2020, also this journal) have recently published a critical response in which they list a number of differences between our two studies which might account for the negative findings. We will use this series of three papers as our exemplum. We also take the opportunity to acknowledge some points of critique provided by Ruiz et al., and to set the record straight with respect to the differences between the original study and our replication attempt. We hope this discussion might help the CBS community to develop a coherent approach to the very current issue of replication.

Keywords: Metaphor; Relational Frame Theory; Replication; Scope; Moderator effects; Falsification

In search of scope: A response to Ruiz et al. (2020)

Replicability is a core feature of science. So much is this the case that the last decade's revelations over failed replications in psychology caused serious questions about the discipline's scientific status (Earp & Trafimow, 2015; Open Science Collaboration, 2015). With this in mind, we set out to replicate an interesting and potentially extremely useful finding — that the inclusion of some fairly simple aspects of language have the power to increase the effect that a metaphor might have on a person's ability to withstand pain in order to do what they really value. Therapists and coaches intervene mostly through language. As a result, Relational Frame Theory (Hayes et al., 2001) — the leading modern behavior analytic theory of language — has been promoted as a framework for improving the effectiveness of such language-based interventions (Foody et al., 2014; McEnteggart et al., 2015; Villatte et al., 2016). The study we aimed to replicate comes out of that tradition. Sierra and colleagues (2016) report an experimental analogue study conducted in a laboratory setting. They test the effect of simple linguistic changes — the inclusion of common physical properties (formal similarities between the metaphorical language and the aversive experience, such as the word "cold" used in a metaphor used before the cold pressor behavioral task) and the evocation of appetitive augmentals ("rules that specify a conditional or causal relation between behaving in a particular way and obtaining abstract positive reinforcers"; Sierra et al., 2016, p. 267) — on pain tolerance. Many scholars consider values, as usually defined in ACT, to function as augmentals (e.g. Levin et al., 2016). Sierra and colleagues (2016) thus operationalized appetitive augmentals by asking participants to imagine something they value as part of the metaphor: "On the other side of the swamp, *there is the most important thing for you, this thing you dream about, the one that excites you the most and makes you vibrate*" (Sierra et al., 2016, p. 271, emphasis in the original). They

find that these changes increase the effect of a metaphor-based verbal intervention and in their conclusion they are clear that they expect the effect they describe to be of broad scope: “according to the results of this study, the ACT therapist from the example of the person in a painful rehabilitation process after surgery should design a metaphor that includes common physical properties with the experienced pain and specify appetitive augmentals to tolerate it” (Sierra et al., 2016, p. 277). If coaches and therapists can increase the power of their verbal interventions by making these sorts of simple linguistic changes, it is clearly of considerable interest. As a team which includes two Coaching Psychologists, we hope that we can help demonstrate that this effect is robust and general.

Having witnessed a number of critiques of related work at the Association for Contextual Behavioral Science Word Conference, across a number of years, we decided not only to replicate Sierra et al. (2016) but also to make what we saw as the most minimal improvements to experimental design that would address these critiques. These critiques, for instance, have referenced differences in intervention script length and the number of incidental analogies (analogies not intended to be linked to the main experimental metaphor) between experimental conditions. We thus described our work as an ‘extended direct replication’ and reported that we had failed to elicit the reported effects (Pendrous et al., 2020). The original authors subsequently published a critique of our replication (Ruiz et al., 2020) in which they raise a number of concerns with how we described our work, and in which they dedicate the majority of their manuscript to describing a series of differences between their original work and our replication attempt.

In this paper, we attempt to address a number of the concerns Ruiz et al. (2020) present. We are happy to join Ruiz and colleagues in setting the record straight where our original

phrasing may have left something to be desired. However, the main purpose of this paper is to describe a significant challenge for contextual behavioral science in dealing with failed replications. We will use the response to our work published by Ruiz et al. (2020) to illustrate the point, but the point is general: it is unwise to attempt to explain away negative findings by pointing to potential moderating variables without considering the implications for the *scope* of our substantive theory. We refer to Hayes et al.'s (2012) definition of scope, that "a given analytic concept applies to a range of cases" (p. 2). This will be expanded upon below.

First, however, we would like to acknowledge the hard work and obvious good intentions of the original authors. Given the importance of metaphors in applied psychological work (and in human discourse generally), it has been extremely encouraging to see the rapidity with which a coherent RFT analysis of metaphorical language has developed. The authors (Ruiz, Luciano, Sierra, Flórez, and Hernández) have played no small part in that important work and we are extremely grateful to them. Indeed, our attempt to replicate this work was not done out of skepticism for their findings, but rather to build on and strengthen what we consider to be excellent work.

Norms of reporting replication work

Ruiz et al. (2020) take issue with our having used the term 'direct replication' when we also state quite plainly that we made alterations to their method. They assert that 'direct replication' is usually reserved for "the repetition of a study to as exact degree as possible" (Ruiz et al., 2020, p. 39). An author's job is to convey their meaning to the reader, and if the use of 'direct' here fails to capture our meaning, then we are only too happy to issue a *mea culpa* and adopt different jargon in future work. Indeed, we might argue that such categories reflect a false binary and might choose simply to avoid such qualifiers in future. That said, it has been argued

that “a direct replication does not have to duplicate all aspects of an original study. Rather it must only duplicate those elements that are believed necessary for producing the original effect.” (Zwaan et al., 2018, pp. 8-9). It is therefore important to consider the nature of the theory under examination in deciding what qualifies as a more or less direct replication.

In replicating a drug trial, the color of the patient information leaflet is unlikely to be of great importance, whilst in a study of attention paid to information leaflets, that same feature may be central to generating the effect. In the former, the theories under test (e.g. that a certain molecule disrupts the cell wall of a particular bacterium, that it has no other deleterious effects on the human body etc.) have nothing to do with the color of the leaflet. In the latter, leaflet color may be a key component of the theory (e.g. that red text garners more attention). We agree with Ruiz et al. that those conducting a *direct* replication, ought to operationalize the key features of a theory in as like a manner to the original study as is practical. However, this process depends not only on the diligence of the replication team. If features of the experimental preparation are not labelled clearly as being necessary to bring about the effect, then no replication team is likely to be able to replicate the effect as they will simply not know what the crucial components are. This holds too for interventionists who seek to apply the theory; without a thorough description of the necessary and sufficient conditions, the interventionist may well fail to bring about the effect. Whilst good interventionists will of course select tools and processes which appear to work for their particular client, it is undeniable that many do use exercises and metaphors in a manner faithful to those presented in papers and textbooks. With only four to six thousand words in most scientific papers, it is unreasonable to expect a research team to describe every conceivable detail of the theory tested. Biochemistry papers do not routinely describe how to apply agar to a petri dish. However, if there is even a reasonable chance that other research teams in the same field

might not know of the possible importance of a given design feature, then it is incumbent upon scientists to describe this feature in their paper, and to describe it as a potentially necessary condition of the observed effect. In the words of Earp and Trafimow (2015, p. 9), "The original investigator should be able to describe exactly what parameters she sees as being theoretically relevant, and under what conditions her "effect" should obtain." Failed replication attempts are not an indication that the original empirical work was flawed. Rather, such failures 'will suggest that the effect is sensitive to theoretically-unspecified factors' (Earp & Trafimow, 2015).

Furthermore, if authors do not label certain features as being important to the effect, then a replication can only be downgraded from 'direct' — because it changes some of these previously undescribed features — after a process of dialogue between teams. To use the response of Ruiz et al. to our study as an example of this: they point out that the sex ratio is different between the two studies, yet their original explication of the effect makes no reference to sex as a moderator. They claim that there were differences in participants' knowledge of ACT/RFT between the two samples, and yet so far as we can see, this was not stated as an inclusion/exclusion criterion in their original study (indeed, they only disclose this in their more recent response to our paper). Similarly, they point out a difference in prior experience with a cold pressor task between our two samples, but at no point in their original paper do they qualify the effect as *only* applicable to naïve participants. They refer to the importance of pauses during the scripts, to permit relational elaboration, yet an explanation for the importance of this feature was missing in their original paper. In light of their painstaking comparison, we are very happy to agree with Ruiz et al. that our replication would better be described as 'conceptual' rather than 'direct'. There is perhaps something for us all to reflect on here, as a scientific community.

Direct replication work requires a number of conditions be met by both the original *and* the replicating team.

We agree with Ruiz et al. that scientists need a better way to describe systematically the differences between initial demonstrations of effect and subsequent replications. As we have indicated, we disagree with their placing the burden for this effort entirely on those conducting replications. In their critique of our replication, Ruiz et al. tabulate 13 differences between our two studies (their Table 1). Their intention appears to suggest that some of these differences might be moderators of the effect. To be clear, a moderator is any variable which "partitions a focal independent variable into subgroups that establish its domains of maximal effectiveness in regard to a given dependent variable" (Baron & Kenny, 1986, p. 1173). We agree that these differences may well account for the difference in results between our studies, however, of these 13 differences, 9 are not mentioned in the original paper as potential moderators of the effect on the independent variable, whilst some others are implied and not stated explicitly. It might be helpful to develop a cultural norm such that all contextual behavioral scientists routinely state any potential moderators of the effect (of anything more than a negligible effect size) of which they are aware, even if those moderators are not under investigation in specific pieces of work.

Replicability has been considered a core feature of science for centuries (Simons, 2014), however, it has typically been difficult to publish either successful or unsuccessful replication studies (Martin & Clarke, 2017). As a team, we have had first-hand experience of rejection owing to 'lack of novelty'. Such norms in the culture of science set up contingencies for the individual scientist. To return to our own example, we sought to 'extend' and improve upon Sierra et al. (2016) partly because we had little confidence that a perfect direct replication would be publishable. Our science would be well served if we can, together, develop a means of

changing the contingencies under which individual scientists are working. Some have argued that replication by the original team is sufficient or even preferable (Cesario, 2014). Such a position would be the very opposite of a CBS-consistent view. Skinner described science as "a corpus of rules for effective action" (Skinner, 1974, p. 235) and recognized that it is a community activity. Verbal rules can guide the action of a single human — the same human from whose experience the rules were derived. However, in the context of the scientific endeavor, the adequacy of such rules can only be tested through the systematic communication of those rules between groups of people who might use the rules to guide action, and through the attempts to use those rules to bring about a similar effect. To borrow Skinner's words, "a proposition is 'true' to the extent that with its help the listener responds effectively to the situation it describes" (Skinner, 1974, p. 235).

Methodological improvements in our replication

Though it is not the main thrust of this paper, we should take the opportunity to put on record a clear account of what we now consider to be the methodological improvements in our study. We attempted (a) to balance the lengths of scripts to address the common 'dose effect' critique of such experiments wherein the addition of linguistic features increases script length for some conditions, (b) to reduce the risk of unblinding and thereby minimize potential for demand characteristics by automating as much of the procedure as possible, (c) to balance the number of descriptive words (e.g. "filthy", "cold") between conditions, (d) to make use of highly standardized cold pressor equipment, (e) to pre-register our study to reduce researcher degrees of freedom in the analysis (see Wicherts et al., 2016). A clumsy use of language on our part gave the impression that we made other improvements, and we are grateful to Ruiz et al. (2020) for setting the record straight.

In search of scope

In many areas of psychology, the business of replication is conceptualized as a fundamental part of the falsificationist approach to science: Theories should be exposed to more and more stringent tests and ultimately, should be revised on the basis of negative findings (Zwaan et al., 2018). In truth, however, when experiments produce negative results, it is usual for scientists to defend their theory against refutation. This tendency has been described at length by any number of historians and philosophers of science (Grünbaum, 1976; Lakatos, 1970). For example, both Meehl (1978, p.819) and Hayes (2004) have described a 'protective belt' of auxiliary assumptions around substantive theories. When negative results are returned, scientists routinely turn on their auxiliaries, most often adding hypothetical moderators of the effect (Grünbaum, 1976). Only when a significant body of negative results builds up, or if auxiliaries and substantive theories are closely connected, are we likely to see a rush to reject the substantive theory (Meehl, 1978, Zwaan et al., 2018).

The above is a rough outline of how a sophisticated falsificationist might view the progress of science. Contextual Behavioral Science, on the other hand, has long seen the problems with this conceptualization of the scientific endeavor (Hayes, 2004) and has instead insisted that the "goal of science is the construction of increasingly organized systems of verbal rules that allow analytic goals to be accomplished with precision, scope, and depth, and based on verifiable experience" (Hayes, 2004, p. 36; see also Hayes, 1993). Within such a framework, how might we conceptualize the usual practice of scientists, searching for potential moderator effects, to explain negative findings? The process maps almost perfectly onto the concept of scope. In the words of Biglan and Hayes (2016, p. 43), "Scope means an analysis is relevant to a broad range of phenomena. For example, the concept of reinforcement has been shown to

account for changes in behavior of humans and nonhumans across an extraordinarily broad range of behaviors." We can readily see, therefore, how introducing hypothetical moderators directly reduces the number of phenomena to which to a theory applies, and thus reduces its scope. Let us first illustrate with a hypothetical example: imagine that we have a general theory that human beings can be taught to accomplish some particular task with their nondominant hand. We run an initial test of the theory and obtain a flattering *p* value. A second team replicates our experiment and the result is negative. We might immediately notice that our sample had more women than men. We therefore hypothesize that sex may be a moderator of our effect: perhaps women are able to learn to accomplish the task with their nondominant hand, whereas men are not. Perhaps this explains our failed replication. In a single stroke we have halved our scope. Our theory no longer applies to the whole the population, only half of it. This is a simplistic example and real cases are rarely so clear cut, but hopefully this example will help even unfamiliar readers to abstract the general principle. Some of the greatest scientists in history seem to have fallen prey to this process, so much so that it has drawn the attention of historians of science (e.g. Lakatos, 1970).

To return to Ruiz et al. (2020), we believe this is precisely the trap into which they risk falling in their response to our replication study. We are in broad agreement with their approach, to now ask whether any of a series of differences between the two studies might be responsible for the differing results. We too are interested to know the boundary conditions of the effect described so far by Sierra et al. (2016) and Criollo et al. (2018). However, as they tabulate these differences, they seem not to consider the implications for scope if any of their potential moderators are in fact responsible. We would have preferred for Ruiz and colleagues to have described the potential implications for scope, as they are likely to be more knowledgeable about

the relevant theory, however, we will attempt such an analysis. Doing so may be illustrative as to the need for careful consideration regarding which factors may likely be responsible for differing results between replications. It should also be held in mind that *depth* — the degree of integration of theories at different level of analysis — is also important in CBS (Hayes et al., 2012), and as a result, if some of the moderators postulated by Ruiz et al. are in fact relevant, some of the implications would be remarkably broad.

First, Ruiz et al. point to "an overrepresentation of females" (p. 39) in our replication. It is undeniably less than ideal to have an imperfect sex ratio in psychological research; however, we have nowhere found any claim that the learning processes described by RFT are sex-dependent. Nor was sex described as a potential moderator in the original study. If the sex ratio proves to be the moderator which reduced the effect in our study, then the immediate implication would seem to be that appetitive augmentals and common physical properties are less potent when communicated to females. Thinking in terms of theoretical *depth* this would seem to suggest that RFT would need to be revised to take account of differential effects with respect to sex. This seems extremely unlikely to us, and so, barring further evidence on the point, we are inclined to disregard sex as a moderator. It should be noted that since we balanced the sex ratio across conditions, differential effects on the cold pressor by sex are irrelevant.

Second, prior knowledge of ACT/RFT is raised as a potential moderator. Sierra et al. in their original paper, do mention this as a variable, but they do not specify this as a specific exclusion criterion for participation. Had the importance of this feature been raised earlier with respect to participant's responses to metaphors, we would have risked the increased participant burden and gathered much more detailed data. We might have considered a more stringent set of inclusion/exclusion criteria in our study. Of greater concern, however, if the putative effects of

common physical properties and appetitive augmentals evaporate once the listener is basically familiar with the ACT model (note that we say "basically familiar" here because at most our participants would have learned of ACT/RFT in just one or two undergraduate-level lectures, none of them focusing on the RFT analysis of metaphors), then this would have considerable ramifications for client work. Instead of urging ACT therapists working in the field of pain to "design a metaphor that includes common physical properties with the experienced pain and specify appetitive augmentals to tolerate it" they would have to add a proviso that such metaphors are only likely to be effective in the first session with a client, whilst they are still entirely naïve to the ACT model.

Third, Ruiz et al. point out that "there were differences in the pauses prompting for relational elaboration across the experimental conditions" (p. 39). Having pointed out that the pauses mid-way through the metaphor-based script varied in length across our conditions, they assert that "[i]n the absence of these pauses, the differences between the experimental protocols in the replication study might be diluted." (p. 43) Ruiz and colleagues have much more experience than us in running laboratory studies testing the RFT conceptualization of metaphors. If they believe that pauses to allow for relational elaboration are necessary, we have no basis on which to disagree. However, we were unable to find any mention of *relational elaboration*, nor of the importance of pauses, in their original paper. Therapists make use of pauses for various theory-driven reasons (Levitt, 2001). If specific lengths of pauses are necessary to bring about the metaphor-boosting effects of common physical properties, this would add a considerable complication to the usability of the theory in the applied setting, and thereby reduce scope once again.

Fourth, in their original paper, Sierra et al. (2016) hypothesize that "the inclusion of common physical properties might show a higher effect in participants with low and medium levels of analogical reasoning abilities" (p. 276). We responded directly to the original authors' suggestion by introducing a measure of analogical reasoning into our study and pre-registering an analysis to test this as a moderator. In seeking explanation for the failed replication, Ruiz et al. make a different claim, asserting that "the moment in which this assessment is conducted might influence the experimental effects on the main dependent variable" (pp. 11-12). The only study (to our knowledge) in the RFT literature that suggests that an analogical reasoning task may prime a participant (or client) enough to boost their apparent analogical reasoning abilities was reported by Carpentier et al. (2002). The effect was found in 5-year old children who were still learning a normal relational repertoire. That said, we agree with the plausibility of this explanation; given the many priming effects reported across psychology, it seems quite possible that a test of analogical reasoning, given at the right time, might temporarily increase the salience of analogical relations thus altering performance on analogical tasks. Again, Ruiz et al. would be in a better position than we to describe the potential ramifications of such an addition to the theory, but it is our view it would likely change the recommendations one might make to coaches and therapists: from 'include common properties in your metaphors' to 'give clients an analogical reasoning task to complete in your waiting room'.

Listing the differences between the two studies which yielded different results may seem, *prima facie*, an innocent enough activity, and an entirely logical choice. However, we must be careful how wide we cast our net. Each difference which is labelled as potentially having caused the difference in results is in fact a modification of the theory. Falsificationists make a distinction between substantive theories and auxiliary assumptions, where the latter are separate and distinct

from the main theory and serve only to permit the operationalization of the theory test (Meehl, 1978). In a field such as CBS, where theoretical development is in the service of applied work, where theory is generated in order to permit the prediction and influence of behavior with breadth, scope, and depth (Hayes et al., 2012), this distinction is of much less utility. If theory is to guide the behavior of coaches and therapists, these professionals must be alerted to *all the necessary conditions* which facilitate the effect. If we adopt a pragmatic truth criterion (Long, 2013) then the truth of our theory is decided upon whether we have laid out a set of verbal rules which enable us successfully to reach our stated aim. If a set of conditions must be in place for us to achieve our aim, then those conditions are *de facto* a part of our theory and must be communicated a such if the theory is to be true.

It is quite possible, of course, that some of these moderators may not be known in early theory development or during the initial experimental investigations of an effect, and that they only become known as a result of differing results between studies. This speaks to the crucial role of replication in any scientific disciplines. By comparing our replication attempt with their own work, Ruiz et al. (2020) have indeed been able to generate a considerable list of hypothetical moderators, some of which may well have been unknown to them at the time of their original study. Whilst it is possibly desirable, in attempting to understand contradictory results, to create as full a list of potential moderators as possible, it is likely that our science would be driven forward much more rapidly if such lists were also accompanied by an analysis of how likely the authors consider each putative moderator to be and how such a moderator effect, if present, would affect the scope of our theory and its applicability to real world applied settings. Further, so as to not dilute our shared scientific vision of a behavioral science integrated with both practice and other evolutionary sciences, we must all be mindful of any implications

that hypothetical moderators may have on the depth of our theories. Journal editors may have a role to play here, in encouraging more systematic communication of such claims, though it is not clear to us how this might be achieved.

The likely status of identified potential moderators

Though our main objective in this paper is to consider how, as a community, we might better support, conduct, and discuss replication work, and how we might best consider the implications when replication attempts fail, it would be remiss of us not to use the opportunity to put on record any further information we may have regarding the potential moderators Ruiz et al. have recently raised. This may help us to work together as a community to discover which of these potential moderators have a noteworthy effect.

Regarding ACT/RFT knowledge: The majority of our sample were students. Some were psychology students. Given the pattern of teaching in our School of Psychology, and the structure of our incentivized participant pool, it is likely that most of those who responded positively to our question about ACT/RFT knowledge were referring to having attended one or two undergraduate lectures on ACT, in which RFT is discussed for approximately 5 minutes. These sessions were not experiential and made no reference to the RFT analysis of metaphor. With hindsight, we perhaps ought to have collected more detailed information on this point, however, we are doubtful that this level of prior experience has a notable effect. Indeed, in response to Ruiz et al. (2020), we re-ran our main analyses on pain tolerance as per our preregistration for the replication study but with ACT/RFT-knowledgeable participants excluded ($N = 50$). There were some increases in effect sizes compared to the original study, though none of the non-significant findings in our original analysis became significant once these participants had been excluded. This is also consistent with research which demonstrates that underpowered

samples can have inflated effects (see 'winner's curse'; Button et al., 2013). (Our re-analyses are available at https://osf.io/p2hwv/).

Ruiz et al. dedicate three paragraphs to a discussion of the differences between our circulating water bath, purchased from a lab supply company, and their artisanal one. They state that "in Sierra et al.'s study, the temperature of the cold pressor task was set at 4.5 to 5.5°C, whereas Pendrous et al. set it at 3°C" (p. 42). This was one of the reasons we claimed to have applied "more stringent conditions"; audits of our cold pressor machine showed that temperature was stable to about 0.1°C whilst in use. They rightly point out (with support from von Baeyer et al., 2011) that cold pressor studies sometimes yield different results across labs, owing to different equipment inducing different levels of pain. The original authors state that their "higher temperature facilitates the use of the strategies trained by the experimenter" (p. 269), but no putative mechanism is described. We originally interpreted this to mean that the participants would be engaged with the cold pressor task long enough to use the techniques and so we piloted our cold pressor set-up to obtain a mean post-intervention cold pressor time in excess of 1 minute. We did not witness sizeable floor or ceiling effects on the cold pressor (the data are available at https://osf.io/p2hwv/). If, instead, Sierra and colleagues meant to suggest that the techniques are only effective with mild discomfort, as induced by a higher cold pressor temperature, then this would suggest a further restriction of scope.

That participants (or clients) must be able to comprehend the language in which the intervention is delivered goes without saying. Ruiz et al. (2020) very reasonably point to the heterogeneity of our sample with respect to first language, all our participants whose first language did not match the language of the intervention (English) were nonetheless sufficiently fluent to study in English, the vast majority of them at postgraduate level. We regrettably omitted

this detail from our paper. Whilst it is likely that delivery in a second language would moderate the effectiveness for less fluent speakers, a careful analysis would be needed to compare postgraduate-level fluency with the fluency of the average client, in order to assess any scope-reducing effect of such a moderator. In passing, it is also worth noting that some of the changes we made to the original scripts were necessary because pilot participants reported to us that they did not understand some of the phrases adopted in the Spanish-to-English translation provided by the original authors (for instance, the verb 'to vibrate' is rarely used to mean excitement in British English). This speaks to yet another challenge our research community will face as replication attempts cross linguistic barriers.

To summarize, Ruiz et al. (2020) point to a number of differences between our replication study and their original work, suggesting that these might explain the differences in results. We agree that some of these putative moderators might plausibly account for the effect. It is impossible to say for sure without further empirical work. However, admitting such moderators into a theory would reduce its scope, in some instances dramatically so. We therefore believe it to be a strategic error to reach for moderators of this sort in the first instance, without exploring other reasons the findings of two (or more) studies disagree.

Building a culture for replication

As a scientific community, just as in the wider field of psychology, we are still discovering how we might best support our fellow scientists to ensure the replicability of our findings across time and settings. We were heartened by the fair-minded response to the contradictory evidence presented in our Pendrous et al. (2020) from reviewers and JCBS editorial team. Until very recently, it has been the publishing norm to reject replication studies on

the grounds that they lack novelty (Martin & Clarke, 2017). It is possible that, as a team, we were still responding to these contingencies. We were imprecise in describing why we considered our replication study to be methodologically more robust; we prioritized describing our work as worthy of publication, emphasizing its merits, instead of, as Ruiz et al. suggest, providing a systematic comparison of the original study and the replication.

It seems likely that Ruiz and colleagues were responding to a similar set of contingencies when they set out to tabulate, rather extensively, all the differences between their original study and our replication. As numerous philosophers of science have pointed out, this is usual practice. However, those same philosophers of science have also reasoned that such attempts to defend a theory against refutation by adding or subtracting auxiliary hypotheses is often problematic. A good deal of extant philosophy of science has been dedicated to an attempt to distinguish science from pseudoscience (Lilienfeld et al., 2012). One of the most often described features of pseudoscience is the manner in which theories are changed, again and again, to keep pace with negative findings. It is clear to us that the teams working on the RFT analysis of analogical reasoning are engaged intelligently and seriously in the honest pursuit of science. However, as a field, we may need to consider how we communicate about our research, so as not to be *seen* to be engaging in pseudoscientific practices by the broad audiences — including journalists and the lay public — who now often read scientific articles. If moderators are to be hypothesized to account for differences in findings across studies, we must also provide a hard-headed analysis of the implications of such putative moderators for the scope, and perhaps also the depth, of our theories. We agree with Ruiz et al. (2020) that it may be salutary to develop a set of community standards for reporting replication studies. However, the adoption of any such approach would warrant robust debate. For instance, it would be foolhardy to reduce method variance artificially

in laboratory-based analogue studies, only to make such studies less ecologically valid. To restrict ourselves to ever more 'direct' replications would serve only to reduce the scope of our scientific theories. We would add that it may be even more important to develop community standards for how we respond to negative results. We do not have the space in this manuscript to go into depth regarding the issue of statistical power, but it may well be that we would do better to consider the potential for false positives and false negatives in our statistical testing before we reach for potential moderators as an explanation of differing findings.

Following our failed replication, Ruiz, Luciano, Sierra, our team, and a whole host of others will now be in search of the boundary conditions of the common-physical-properties effect. As we engage in this effort, we cannot admit all possible moderators of the effect into our theory without careful reasoning about what each would mean for the scope, and therefore the usefulness, of our theory. Our shared mission is to create "a science more adequate to the challenge of the human condition" (Hayes et al., 2012). If we are left with a theory which explains the behavior only of males who are entirely naïve to both RFT and ACT, who experience only a very particular level of pain, and whose analogical reasoning abilities are suboptimal, our theories will so lack scope that we will have served no one.

References

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173-1182.

Biglan, A., & Hayes, S. C. (2016) Functional contextualism and Contextual Behavioral Science. In R. D. Zettle., S. C. Hayes., D. Barnes-Holmes, & A. Biglan (Eds.), *The Wiley Handbook of Contextual Behavioral Science* (pp. 37-62). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118489857

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365-376. https://doi.org/10.1038/nrn3475

Carpentier, F., Smeets, P. M., & Barnes-Holmes, D. (2002). Matching functionally same relations: Implications for equivalence-equivalence as a model for analogical reasoning. *The Psychological Record*, *52*(3), 351-370. https://doi.org/10.1007/BF03395435

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives in Psychological Science*, *9*(1), 40-48. https://doi.org/10.1177/1745691613513470

Criollo, A. B., Díaz-Muelle, S., Ruiz, F. J., & García-Martín, M. B. (2018). Common physical properties improve metaphor effect even in the context of multiple examples. *The Psychological Record*, *68*(4), 513-523. https://doi.org/10.1007/s40732-018-0297-9

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*(621), 1-11. https://doi.org/10.3389/fpsyg.2015.00621

Foody, M., Foody, M., Barnes-Holmes, Y., Barnes-Holmes, D., Törneke, N., Luciano, C., Luciano, C., Stewart, I., McEnteggart, C., & McEnteggart, C. (2014). RFT for clinical use: The example of metaphor. *Journal of Contextual Behavioral Science*, *3*(4), 305-313. https://doi.org/10.1016/j.jcbs.2014.08.001

Grünbaum, A. (1976). Ad hoc auxiliary hypotheses and falsificationism. *The British Journal for the Philosophy of Science*, *27*(4), 329-362. https://doi.org/10.2307/686862

Hayes, S. C. (1993). *Analytic Goals and the Varieties of Scientific Contextualism*. In S. C. Hayes, L. J. Hayes, H. W. Reese, & T. R. Sarbin (Eds.), *Varieties of Scientific Contextualism* (pp. 11-27). Context Press.

Hayes, S. C. (2004). Falsification and the protective belt surrounding entity-postulating theories. *Applied and Preventive Psychology*, *11*(1), 35-37. https://doi.org/10.1016/j.appsy.2004.02.004

Hayes, S. C., Barnes-Holmes, D., & Wilson, K. G. (2012). Contextual Behavioral Science: Creating a science more adequate to the challenge of the human condition. *Journal of Contextual Behavioral Science*, *1*(1-2), 1-16. https://doi.org/10.1016/j.jcbs.2012.09.004

Hayes, S. C., Barnes-Holmes, D., & Roche, B. (2001). *Relational Frame Theory: A post-Skinnerian account of human language and cognition*. Kluwer Academic Publishers.

Lakatos, I. (1970). Falsification and the Methodology of Scientific Research Programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge: Proceedings of the International Colloquium in the Philosophy of Science, London, 1965* (pp. 91-196). Cambridge University Press.

Levin, M. E., Twohig, M. P., & Smith, B. M. (2016) Contextual Behavioral Science: An Overview. In R. D. Zettle., S. C. Hayes., D. Barnes-Holmes, & A. Biglan (Eds.), *The*

Wiley Handbook of Contextual Behavioral Science (pp. 17-36). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118489857

Levitt, H. M. (2001). Sounds of silence in psychotherapy: The categorization of clients' pauses. *Psychotherapy Research*, *11*(3), 295-309. https://doi.org/10.1080/713663985

Lilienfeld, S. O., Ammirati, R., & David, M. (2012). Distinguishing science from pseudoscience in school psychology: Science and scientific thinking as safeguards against human error. *Journal of School Psychology*, *50*(1), 7-36. https://doi.org/10.1016/j.jsp.2011.09.006

Long, D. M. (2013). Pragmatism, realism, and psychology: Understanding theory selection criteria. *Journal of Contextual Behavioral Science*, *2*(3-4), 61-67. https://doi.org/10.1016/j.jcbs.2013.09.003

Martin, G. N., & Clarke, R. M. (2017). Are psychology journals anti-replication? A snapshot of editorial practices. *Frontiers in Psychology*, *8*(523), 1-6. https://doi.org/10.3389/fpsyg.2017.00523

McEnteggart, C., Barnes-Holmes, Y., Hussey, I., & Barnes-Holmes, D. (2015). The ties between a basic Science of language and cognition and clinical applications. *Current Opinion in Psychology*, *2*, 1-4. https://doi.org/10.1016/j.copsyc.2014.11.017

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806-834.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716-aac4716. https://doi.org/10.1126/science.aac4716

Pendrous, R., Hulbert-Williams, L., Hochard, K. D., & Hulbert-Williams, N. J. (2020). Appetitive augmental functions and common physical properties in a pain-tolerance

metaphor: An extended replication. *Journal of Contextual Behavioral Science*, *16*, 17-24. https://doi.org/10.1016/j.jcbs.2020.02.003

Ruiz, F. J., Luciano, C., & Sierra, M. A. (2020). A systematic and critical response to Pendrous et al. (2020) replication study. *Journal of Contextual Behavioral Science, 17*, 39-45. https://doi.org/10.1016/j.jcbs.2020.04.011

Sierra, M. A., Ruiz, F. J., & Flórez, C. L. (2016). The role of common physical properties and augmental functions in metaphor effect. *International Journal of Psychology and Psychological Therapy*, *16*(3), 265-279.

Simons, D. J. (2014). The value of direct replication. *Perspectives in Psychological Science*, *9*(1), 76-80. https://doi.org/10.1177/1745691613514755

Skinner, B. F. (1974). *About Behaviorism*. Vintage Books.

Villatte, M., Villatte, J. L., & Hayes, S. C. (2016). *Mastering the Clinical Conversation: Language as Intervention*. The Guilford Press.

von Baeyer, C. L., Torvi, D., Hemingson, H., & Beriault, D. (2011). Water circulation and turbulence in the cold pressor task: Unexplored sources of variance among experimental pain laboratories. *Pediatric Pain Letter*, *13*(1), 13-16.

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., van Aert, R. C., & van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*(1832), 1-12. https://doi.org/10.3389/fpsyg.2016.01832

Zwaan, R., Etz, A., Lucas, R., & Donnellan, M. (2018). Making replication mainstream. *Behavioral and Brain Sciences, 41*, E120. https://doi.org/10.1017/S0140525X17001972