MANUSCRIPT PUBLISHED IN JOURNAL OF CONTEXTUAL BEHAVIORAL SCIENCE

A systematic and critical response to Pendrous et al. (2020) replication study

Francisco J. Ruiz[1]

Carmen Luciano[2,3]

Marco A. Sierra[4]

[1]Fundación Universitaria Konrad Lorenz, [2]Universidad de Almería, [3]Madrid Institute of Contextual Psychology, [4]Universidad Católica de Colombia,

Correspondence address: Francisco J. Ruiz, franciscoj.ruizj@konradlorenz.edu.co, Fundación Universitaria Konrad Lorenz, Carrera 9 bis, Nº 62-43, Bogotá (Cundinamarca, Colombia), Phone: (+57 1) 347 23 11 ext. 185

Abstract

Conducting direct replication studies is crucial for the progress of science because they increase our confidence in the effect of the independent variables under the same or mostly the same experimental conditions. Pendrous et al. (2020) recently published an "extended direct replication" with negative results concerning the study by Sierra et al. (2016) and suggested the disparity in results was due to the supposed more stringent conditions of their study. However, a detailed comparison of the studies reveals that: (a) they differed in many relevant aspects (e.g., participants' characteristics, experimental task, procedure, and experimental protocols) that preclude considering Pendrous et al.'s study as a "direct replication;" (b) the replication study did not specify some methodological strengths of the original study; and (c) the replication study had unnoticed methodological problems. In the replication study: (a) there was an overrepresentation of females, (b) there were notable differences across experimental conditions in the naïve status of the participants in terms of previous ACT/RFT knowledge and experience with the cold pressor task, (c) 21.4% of the participants were not native English speakers, (d) compensation was not the same for all participants, and (e) there were differences in the pauses prompting for relational elaboration across the experimental conditions. These methodological problems might limit the conclusions reached in the replication study. We call for greater precision in reporting and discussing replication studies by highlighting the commonalities and differences between the original and replication studies.

Key words: Replication crisis; Metaphor; Psychological flexibility; Relational frame theory; Acceptance and commitment therapy.

A systematic and critical response to Pendrous et al. (2020) replication study

1. The relevance of replication studies and their systematic reporting

The claim for replication as a pivotal strategy has been generalized in the history of science in general and psychology in particular (e.g., Gasparikova-Krasnec & Ging, 1987; Johnston & Pennypacker, 1980; Rosenthal, 1990; Sidman, 1960). A paradigmatic case in this regard constitutes the behavior analytic tradition in which replication was defined as the core criterion for the inductive formulation of principles of behavior (e.g., Ferster & Skinner, 1957; Johnston & Pennypacker, 1980; Sidman, 1960; Skinner, 1938). Two types of replication studies were soon differentiated: direct and systematic replications (Hersen & Barlow, 1976; Kazdin, 1982; Sidman, 1960).

A direct replication involves the repetition of a study to as exact degree as possible, intending to identify if the conditions under which the effect was first demonstrated produce it reliably. Direct replications are essential to increase our confidence in the effect of the independent variables under the same experimental conditions.

A systematic replication involves the repetition of a study by changing one or more conditions to identify the generality of the previous findings. If systematic replications find similar effects under different experimental conditions, our confidence in the generality of the findings increases.

Conducting and publishing both direct and systematic replications is necessary for the progress of science. However, replication studies have historically experienced difficulties in being published for at least two reasons (e.g., Makel, Plucker, & Hegarty, 2012): (a) the lack of interest due to the low originality of the studies that find the same pattern of results as the original study, and (b) because results supporting the null

hypothesis, when not replicating the initial results, used to be rejected. Fortunately, there is a recent changing trend with the so-called replication crisis that is motivating numerous journals to accept replication studies (e.g., Nosek et al., 2015). This is the context for the current article.

In this article, we highlight that the reporting of replication studies should systematically compare the studies in question (e.g., Brandt et al., 2014; Johnston & Pennypacker, 1980). This systematic comparison is even more critical when a disparity in results is observed. If the studies are not systematically compared, the scientific community and the discipline might become saturated with original and replication studies that led to different results with the frustration of not knowing, or at least suspecting, under which conditions experimental manipulations had an effect.

Recently, Pendrous, Hulbert-Williams, Hochard, and Hulbert-Williams (2020) published "an extended direct replication" of the study conducted by Sierra, Ruiz, Flórez, Riaño-Hernández, and Luciano (2016), which was also partly replicated by Criollo, Díaz-Muelle, Ruiz, and García-Martín (2018). The replication study found different results from the original study and claimed that "However, given that the current study had many methodological strengths in comparison to these two studies (e.g., double-blind randomization, stratification), suggesting that under more stringent experimental conditions, we do not yet have sufficient knowledge of the contextual factors affecting the power of metaphors to change behavior" (p. 23). Also, on the same page, the same idea was repeated: "Under more stringent conditions, the key hypothesis was not supported as pain tolerance did not increase after any metaphor intervention." (p. 23).

Is the claim of more stringent experimental control justified? What are the similarities and fundamental differences between the original and replication study? Can

Pendrous et al.'s study be considered a direct replication of Sierra et al.'s? Could the differences found between the studies help to suspect why there was a divergence in the results? In this article, we aim to respond to these questions systematically and critically.

2. The original study by Sierra et al. (2016)

Previous studies based on relational frame theory (RFT; Hayes, Barnes-Holmes, & Roche, 2001) and acceptance and commitment therapy (ACT; Hayes, Strosahl, & Wilson, 1999; Wilson & Luciano, 2002) served as an inspiration for Sierra et al. (2016). On the one hand, several studies showed that including values in analogs of therapy interventions increased their effect on pain tolerance (e.g., Branstetter-Rost, Cushing, & Douleh, 2009; Gutiérrez, Luciano, Rodríguez, & Fink, 2004; Luciano et al., 2014; Páez-Blarrina et al., 2008). On the other hand, in a basic RFT study, Ruiz and Luciano (2015) showed that relating relational networks (i.e., analogy) with common physical properties was considered as more apt than relating networks without them. In other words, common physical properties increased analogy aptness.

Sierra et al. (2016) analyzed the role of two variables in the effect of metaphors in increasing pain tolerance in a cold pressor task: (a) the presence of common physical properties between the individual's pain and the metaphor, and (b) the specification of appetitive augmental functions (i.e., values) in the metaphor content. In Phase 1, eighty-three participants who were naïve in terms of knowledge about ACT/RFT and experience with the cold pressor task provided informed consent. Afterward, they responded to measures of experiential avoidance using the Acceptance and Action Questionnaire (AAQ-II; Bond et al., 2011; Ruiz et al., 2016), cognitive fusion using the Cognitive Fusion Questionnaire (CFQ; Gillanders et al., 2014; Ruiz, Suárez-Falcón, Riaño-Hernández, &

Gillanders, 2017), and generalized pliance using the Generalized Pliance Questionnaire (GPQ; Ruiz, Suárez-Falcón, Barbero-Rubio, & Flórez, 2019). In Phase 2, participants were exposed to a cold pressor task as a pretest. The three participants who tolerated the pain for 300 s in this phase were excluded from further participation in the study because they reached the maximum time for ethical reasons.

In Phase 3, the remaining 80 participants (40 females) were randomly assigned to four experimental conditions and listened to a protocol displaying the swamp metaphor (Gutiérrez et al., 2004; Hayes et al., 1999; Wilson & Luciano, 2002). In Condition A, the metaphor contained common physical properties with the participants' experience (i.e., the water of the swamp was very cold, as it was the water in the cold pressor task) and appetitive augmentals (i.e., on the other side of the swamp was the most important thing for the participant). In Condition B, the metaphor only contained common physical properties (i.e., on the other side of the swamp was a landscape similar to the participant's side). In Condition C, the metaphor only contained appetitive augmentals (i.e., the water of the swamp was filthy). Lastly, in Condition D, the metaphor did not include these variables (i.e., the water was filthy, and there was the same landscape on the other side of the swamp). In Phase 4, participants were re-exposed to the cold pressor task as a posttest.

The results showed that both independent variables (i.e., common physical properties and appetitive augmentals) had a significant effect on pain tolerance. Accordingly, the authors suggested ACT therapists that explicitly including both components of the metaphor might increase their efficacy.

3. A systematic comparison of Sierra et al. (2016) and Pendrous et al. (2020)

Pendrous et al. (2020) stated that their study is an extended direct replication of Sierra et al. (2016). Contrarily to the original study, the replication study did not find statistically significant differences across their experimental conditions. To explain these different results, the replication study claimed they might be due to the more stringent conditions under which the replication study was conducted. However, we have three critical concerns regarding these statements.

The first concern is that there were important differences between the studies in terms of the participants' characteristics, experimental task, procedure, and experimental protocols. Unfortunately, these differences were not systematically described in the replication study, which makes it difficult to conclude that the authors did not conduct a direct replication study. As Pendrous et al.'s study was not a direct replication, it is difficult to reach strong conclusions when comparing the results of both studies as the authors did. The differences in the results should be discussed taking into account the differences between the studies instead of making statements based on the supposed methodological superiority of the replication study.

The second concern is that the replication study did not describe some relevant methodological strengths of the original study, as we will demonstrate below. Consequently, the conclusion that the divergence in results might be due to the more stringent experimental control exerted in the replication study is even less justified.

The third concern is that a detailed inspection of the replication study reveals that there were important differences across their experimental conditions that might minimize the internal validity of the replication study. As we will demonstrate below, these differences limit the conclusion reached in the replication study.

Next, we are going to compare both studies systematically. A summary of this comparison is presented in Table 1.

INSERT TABLE 1 ABOUT HERE

3.1. Randomization and blinding

The replication study stated that they made some changes to test the reproducibility of Sierra et al.'s results under more stringent conditions: "First, the computerized software PsychoPy (Peirce et al., 2019) was used to automate the task. This was in order to reduce potential experimenter effects and facilitate a truly double-blind design. Second, given sex differences often reported in studies using the cold-pressor task (Mitchell, MacDonald, & Brodie, 2004), participants were randomized by condition, but the ratio of males and females balanced across all conditions (p. 18)." In other parts of the article, it is said that "We tested the replicability of these findings under more stringent conditions, using a stratified (by sex) double-blind randomized-controlled experimental design" (p. 17).

The statements made in the replication study implied that Sierra et al.'s study: (a) did not randomly assign participants to the experimental conditions, (b) did not blind the experimenter, and (c) that the ratio of males and females were not balanced across conditions. However, the authors did not note that Sierra et al. stated the following regarding randomization and stratification by sex: "Participants were randomly allocated to the experimental conditions with the sole restriction of maintaining the same proportion of men and women because previous research has shown some gender differences in performance of the cold-pressor task (e.g., Keogh et al., 2005; Pokhrel et al., 2013)" (p. 269). Interestingly, while Sierra et al.'s study had the same percentage of male and female participants, in the replication study 77.5% of participants were females (p. 21).

Additionally, the replication study did not note that the experimenter in Sierra et al.'s study was blinded as it was stated in the sentence: "To avoid the potential influence of the experimenter's expectations, he did not know which protocol the participant was hearing" (p. 272).

In conclusion, Pendrous et al.'s study repeated several times that their study was conducted under more stringent conditions by "using a stratified (by sex) double-blind randomized-controlled experimental design." However, Sierra et al. (2016) also followed this experimental design with the addition of having the same percentage of males and females. In contrast, the proportion of females was biased in the replication study. Thus, their claim of more stringent methodological conditions does not seem to be justified.

Above all the previous details, the critical point is that the comparison of the results of both studies should have taken into account the difference in sex representation instead of incorrectly highlighting that the replication study was conducted under more stringent conditions.

3.2. Participants' characteristics

Participants in Sierra et al. (2016) were naïve in terms of knowledge about ACT/RFT and experience with the cold pressor task (p. 268). However, 40.5% and 23.8% of the participants in the replication study reported having at least some knowledge of ACT/RFT and previous experience with the cold pressor task, respectively (p. 21). Importantly, when inspecting the dataset shared in the replication study at https://osf.io/p2hwv/, we found statistically significant differences across conditions regarding knowledge of ACT/RFT (see Table 2). Regarding the previous experience with

the cold pressor task, there were marginally statistical differences across conditions (see Table 3).

INSERT TABLE 2 ABOUT HERE

INSERT TABLE 3 ABOUT HERE

The inspection of the dataset also reveals notable differences in the naïve status of participants regarding knowledge of ACT/RFT and experience on the cold pressor (see Table 4). For instance, whereas almost all participants in Condition C were naïve (90.5%), only 20% of participants in Condition B were naïve.

INSERT TABLE 4 ABOUT HERE

In Sierra et al. (2016), all participants were undergraduates, and participation was compensated in the same way (p. 269); however, in the replication study participants were undergraduates and members of the university staff, and they were compensated in different ways (p. 19). Additionally, all participants in the original study had the same first language (i.e., Spanish); however, 78.6% of participants had the same first language (i.e., English) in the replication study (p. 21).

In conclusion, Pendrous et al.'s study incorporated extraneous variables that were not present in the original study: (a) 40.5% of participants had at least some knowledge of ACT/RFT and there were statistically significant differences in the distribution of these participants, (b) 23.8% of participants had previous experience with the experimental task, (c) there were notable differences across conditions in the naïve status of participants regarding the first two points, (d) participants were compensated in different ways, and (e) 21.4% of the participants were not native English speakers.

As indicated before, the description of these differences between the studies would have been needed to take them into account when discussing the results found in the replication study.

3.3. Experimental task

Both studies used the cold pressor task to test the effects of the independent variables. As in many other studies in Psychology, Sierra et al. (2016) used an artisanal apparatus: "… a 30x20x20 cm glass container with two interconnected compartments: one for the ice and the other for the water. In the latter compartment, participants introduced their hands. A digital thermometer was adhered to the container to control the water temperature. Two water pumps (300 liters per hour) were also adhered to the glass container to maintain the water circulating. An ice maker machine was used to keep the temperature of the glass container constant" (p. 269). Pendrous et al. (2020) used an automatic cold pressor: "The cold pressor machine (Jeio Tech Refrigerating Bath Circulator RW-2025G) provides a 15×20 inch basin filled water and circulated at a constant temperature of 3°C" (p. 19).

Note that in both studies, the water was circulating, which is important to avoid the formation of heat bags around the hand (e.g., von Baeyer, Torvi, Hemingson, & Beriault, 2011). However, 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.

In conclusion, the studies used different apparatus and temperatures. There is empirical evidence that even slight changes in temperatures can lead to different results (von Baeyer et al., 2011). Thus, according to the literature on the cold pressor task, the difference in temperature between the studies might have been very relevant. Sierra et al.

(2016) justified their temperature this way: “According to Mitchell, MacDonald, and Brodie (2004), this higher temperature facilitates the use of the strategies trained by the experimenter” (p. 269). However, the replication study did not report why they set the temperature at 3 °C, which could have affected the potential efficacy of the experimental protocols.

3.4. Procedure

Firstly, in Sierra et al. (2016), the administration of the experimental procedure was conducted in the presence of the experimenter. However, in the replication study, the procedure was administered through PsychoPy (p. 19), and it is not clear if the experimenter was present during the cold pressor task. If that was the case, the procedure administration was very similar across the studies, with the main difference of responding on paper or the computer. However, if the experimenter was not in the experimental room during the cold pressor task, that would be a significant difference between the studies (von Baeyer et al., 2011). If the experimenter was present in the room, it would have been important to specify the experimenter’s sex because research has shown that this factor might have a significant effect (e.g., male participants might show higher tolerance in the presence of a female experimenter than with a male one; Vigil, Rowell, Alcock, & Maestes, 2014).

Secondly, one of the reasons that Pendrous et al. (2020) argued to claim that they conducted an “extended direct replication” is that they administered an analogical reasoning test: the Wechsler Adult Intelligence Scale®—Third Edition (WAIS-III; Wechsler, 1997), Similarities Subtest of the Verbal Subscale. We made this suggestion in the original study; however, the moment in which this assessment is conducted might

influence the experimental effects on the main dependent variable. Pendrous et al. administered the analogical reasoning test at the pretest. This might have some unwanted effects because the participants were practicing the analogical repertoire just a few minutes before listening to the experimental protocols. Establishing an explicit context of analogical reasoning in the experimental procedures might have led participants to derive the rule that they had to find the common elements between the cold pressor task and the experimental protocols. This might provoke two undesired effects in terms of experimental control: (a) to help participants to understand less apt analogies, and (b) to focus participants' attention on judging how good the metaphor they were listening was, which could impede contacting with the emotional functions involved in appetitive augmentals.

3.5. Experimental protocols

According to Pendrous et al. (2020), the experimental protocols were modified in the following way: "The original metaphor scripts from Sierra et al. (2016) were edited in the following ways: (i) all conditions were made equal in both word count, audio length, and number of qualifying words (e.g., "very," "awful") and adjectives (e.g., "cold," filthy"), and (ii) some phrases were altered to make the scripts flow more naturally, following a translation from Spanish to English. These were so the conditions were standardized and so the conditions which include no common properties did not include an analogy as descriptions of the swamp ("smells like a sewer") within the metaphor as it is unclear what effect this additional analogical relation would have" (p. 18). However, we observed more differences among the protocols than the ones described in the replication study.

Firstly, the wordings for manipulating appetitive augmentals were considerably different between studies. Specifically, Sierra et al. (2016) introduced appetitive augmentals by saying, "… there is the most important thing for you, this thing you dream about, the one that excites you the most and makes you vibrate…" (p. 271). Pendrous et al.'s wording was "… there is the most important thing to you – something that really excites you, or something that you often think about…" (p. 20). The expression "or something that you often think about" might have been confusing for the participants because it might have actualized the functions of problems they usually ruminate on. This might be a very relevant difference between the original and replication study.

Secondly, Sierra et al. assessed protocol understanding immediately before conducting the posttest (p. 272), but the replication study did not (p. 23). This is a limitation of because it cannot be stated that the independent variables were functionally manipulated. Again, this might be a relevant difference between the original and replication studies.

Thirdly, the protocols designed by Sierra et al. contained several crucial pauses to promote the relational activity of the metaphor (see Table 5). The first pause (15 s) was included to actualize the aversive functions experienced during the first exposure to the cold pressor task so that they could be related in coordination with the aversive functions contained in the metaphors (i.e., crossing the swamp by swimming in cold or filthy water). The second pause (30 s) was included for participants to vividly experience what was on the other side of the swamp (i.e., appetitive augmentals or values vs. the same landscape). The third pause (15 s) aimed for the participants to actualize emotional functions related to advancing towards their values vs. reaching the same landscape. Lastly, the fourth pause

(10 s) was introduced to promote participants' willingness to be in contact with the aversive functions included in swimming towards the other side of the swamp.

INSERT TABLE 5 ABOUT HERE

We accessed the audio files provided by in the replication study at https://osf.io/p2hwv/ and found three important issues. Firstly, there were significant differences between the scripts reported in the article and the content of the audio files. Secondly, Table 5 shows that the duration of the first and fourth pauses was not the same across conditions in the replication study. Lastly, the pauses introduced in the replication study were considerably briefer than those in the original study. The most significant differences between the pauses of both studies were on the second and third ones (see Table 5). In the absence of these pauses, the differences between the experimental protocols in the replication study might be diluted.

In summary, the experimental protocols used in the replication study were considerably different from the ones used in the original study due to: (a) the wording to introduce appetitive augmentals, (b) the absence of an assessment of the protocol understanding, and (c) the smaller pauses to prompt relational elaboration. Additionally, there were differences in the duration of the pauses across the experimental protocols in the replication study.

4. Discussion

As commented above, we have three critical concerns regarding the replication study: (a) the important differences between the original and replication study, (b) the unnoticed methodological strengths of the original study, and (c) the differences in their experimental conditions. The next lines summarize them.

4.1. Important differences between the studies and unnoticed methodological strengths of the original study

We have demonstrated that there were more differences between the studies than the ones indicated in the replication study. Regarding participants' characteristics, the differences were: (a) the ratio of male and female participants was significantly different (50% in the original study vs 77.5% of females in the replication study), (b) the replication study included participants that were not naïve in terms of knowledge in ACT/RFT (40.5%), (c) the replication study included participants with previous experience with the cold pressor task (23.8%), and (d) the replication study included participants who were exposed to the experimental procedures in their second language (21.4%). With respect to the experimental task, the replication study: (a) used a different type of apparatus, and (b) a lower temperature than in the original study. At the procedure level, the replication study: (a) implemented the experiment automatically whereas the original study was implemented with the presence of an experimenter, and (b) included an analogical reasoning test at pretest. Lastly, regarding the experimental protocols: (a) the wordings for manipulating appetitive augmentals were considerably different between studies, (b) protocol understanding was not assessed in the replication study, and (c) the replication study did not introduce the pauses that prompted relational elaboration in the original study.

The abovementioned differences preclude considering Pendrous et al.'s study as a "direct replication." Consequently, the replication study should have discussed the results taking into account the differences between the studies instead of making statements based on the supposed methodological superiority of their study.

Regarding the latter point, we have also demonstrated that Pendrous et al. did not describe some methodological strengths of the original study. Specifically, the replication

study implied that the original one: (a) did not randomly assign participants to the experimental conditions, (b) did not blind the experimenter, and (c) that the ratio of males and females was not balanced across conditions. We have demonstrated with quotes from the original study that Sierra et al. randomly assigned participants to the experimental conditions, blinded the experimenter, and balanced the ratio of males and females across experimental conditions. In conclusion, the claim presented in the replication study of having conducted the study under more stringent methodological conditions is not justified.

4.2. Methodological problems of the replication study.

We have shown that the replication study had methodological deficiencies that went unnoticed in the article. The most important deficiencies are: (a) 40.5% of participants had at least some knowledge of ACT/RFT, with experimental conditions showing statistically significant differences in their distribution; (b) 23.8% of participants had previous experience with the cold pressor task, with marginally significant differences across conditions; (c) there were notable differences in the naïve status of the participants across conditions; (d) 21.4% of the participants were not native English speakers; (e) participants' compensation was not the same for all participants; and (f) the duration of the pauses in the protocols was different across experimental conditions.

In our view, the replication study has some methodological problems that might limit the conclusions reached when comparing the performances of the different experimental conditions.

4.3. Conclusions

Direct replications are needed to confirm that a particular effect is obtained under precise experimental conditions, which include the interaction between the participants'

characteristics and the experimental procedures and protocols. Systematic replications might be conducted afterward to explore to what extent the experimental effect generalizes under different participants and conditions.

Pendrous et al. are to be commended for trying to conduct a direct replication study of the Sierra et al.'s study. However, although we believe that it is a merit to conduct this type of studies and that it involves many hours of hard work, a detailed inspection of the replication study reveals that it was not a direct replication due to the number of important differences between the studies. Unfortunately, the article of the replication study presented the results as a probe of the lack of replication of previous findings (Criollo et al., 2018; Sierra et al., 2016) and repeatedly suggested that it could be due to the more rigorous experimental control they exerted in their study. We have demonstrated that this was not the case.

We hope this article has served the purpose of highlighting the need to compare the original and replication studies systematically. As commented in the introduction, conducting replication studies and systematically reporting them is crucial for the progress of science. The reporting process should make the similarities and differences between the original and replication studies explicit in terms of the manipulation of independent variables and the conditions under which they were implemented. This way, we can avoid creating unnecessary confusion within the discipline and foster a better understanding of the conditions responsible for the observed experimental effects.

We would like to close this article by emphasizing two important points. First, in the era of the so-called replication crisis, we think it is fair to reclaim the role replication has had in this behavioral tradition since its origin as the crucial criterion that permits the inductive formulation of behavior principles (e.g., Ferster & Skinner, 1957; Johnston &

Pennypacker, 1980; Sidman, 1960; Skinner, 1938). Second, we emphasize the crucial importance of reporting replication studies systematically to foster scientific progress and avoid unnecessary confusion in the discipline.

References

Bond, F. W., Hayes, S. C., Baer, R. A., Carpenter, K. M., Guenole, N., Orcutt, H. K., … Zettle, R. D. (2011). Preliminary psychometric properties of the Acceptance and Action Questionnaire – II: A revised measure of psychological inflexibility and experiential avoidance. *Behavior Therapy, 42*, 676-688.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., ... van 't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50*, 217-224.

Branstetter-Rost, A., Cushing, C., & Douleh, T. (2009). Personal values and pain tolerance: Does a values intervention add to acceptance? *The Journal of Pain, 10*, 887-892.

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*, 513-523.

Ferster, C. B., & Skinner, B. F. (1957). *Schedules of reinforcement*. New York: Appleton-Century-Crofts.

Gasparikova-Krasnec, M., & Ging, N. (1987). Experimental replication and professional cooperation. *American Psychologist, 41*, 266-267.

Gillanders, D. T., Bolderston, H., Bond, F. W., Dempster, M., Flaxman, P. E., Campbell, L., … Remington, B. (2014). The development and initial validation of the Cognitive Fusion Questionnaire. *Behavior Therapy, 45*, 83-101.

Gutiérrez, O., Luciano, C., Rodríguez, M., & Fink, B. C. (2004). Comparison between an acceptance-based and a cognitive-control-based protocol for coping with pain. *Behavior Therapy, 35*, 767-783.

Hayes, S. C., Barnes-Holmes, D., & Roche, B. (2001). *Relational frame theory. A post-Skinnerian account of human language and cognition*. New York: Kluwer Academic Press.

Hayes, S. C., Strosahl, K. D., & Wilson, K. G. (1999). *Acceptance and commitment therapy. An experiential approach to behavior change*. New York: Guilford Press.

Hersen, M., & Barlow, D. H. (1976). *Single case experimental designs: Strategies for studying behavior change*. New York: Pergamon Press.

Johnston, J. M., & Pennypacker, H. S. (1980). *Strategies and tactics of behavioral research*. New York: Routledge

Kazdin, A. E. (1982). Single-case experimental designs in clinical research and practice. *New Directions for Methodology of Social & Behavioral Science, 1*3, 33–47.

Keogh, E., Bond, F. W., Hanmer, R., & Tilston, J. (2005). Comparing acceptance-and control-based coping instructions on the cold-pressor pain experiences of healthy men and women. *European Journal of Pain, 9,* 591-591.

Luciano, C., Valdivia-Salas, S., Ruiz, F. J., Rodríguez-Valverde, M., Barnes-Holmes, D., Dougher, M. J., ... & Gutierrez-Martínez, O. (2014). Effects of an acceptance/defusion intervention on experimentally induced generalized avoidance: A laboratory demonstration. *Journal of the Experimental Analysis of Behavior, 101,* 94-111.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science, 7*, 537-542.

Mitchell, L. A., MacDonald, R. A., & Brodie, E. E. (2004). Temperature and the cold pressor test. *The Journal of Pain, 5*, 233-237.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... & Contestabile, M. (2015). Promoting an open research culture. *Science, 348*, 1422-1425.

Páez-Blarrina, M., Luciano, C., Gutiérrez-Martínez, O., Valdivia, S., Ortega, J., & Rodríguez-Valverde, M. (2008). The role of values with personal examples in altering the functions of pain: Comparison between acceptance-based and cognitive-control-based protocols. *Behaviour Research and Therapy, 46*, 84-97.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., ... Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods, 51*, 195-203.

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.

Pokhrel, B. R., Malik, S. L., Ansari, A. H., Paudel, B. J., Sinha, R., & Sinha, M. (2013). Effect of sub-maximal exercise stress on cold pressor pain: A gender based study. *Kathmandu University Medical Journal, 11*, 54-59.

Rosenthal, R. (1990). How are we doing in soft psychology? *American Psychologist, 45*, 775–777.

Ruiz, F. J., & Luciano, C. (2015). Common physical properties among relational networks improve analogy aptness. *Journal of the Experimental Analysis of Behavior, 103*, 498-510.

Ruiz, F. J., Suárez-Falcón, J. C., Barbero-Rubio, A., & Flórez, C. L. (2019). Development and initial validation of the Generalized Pliance Questionnaire. *Journal of Contextual Behavioral Science, 12*, 189-198.

Ruiz, F. J., Suárez-Falcón, J. C., Cárdenas-Sierra, S., Durán, Y., Guerrero, K., & Riaño-Hernández, D. (2016). Psychometric properties of the Acceptance and Action Questionnaire – II in Colombia. *The Psychological Record, 66*, 429-437.

Ruiz, F. J., Suárez-Falcón, J. C., Riaño-Hernández, D., & Gillanders, D. (2017). Psychometric properties of the Cognitive Fusion Questionnaire in Colombia. *Revista Latinoamericana de Psicología, 49*, 80-87.

Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology*. New York: Basic Books.

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

Skinner, B. F. (1938). *The behavior of organisms*. New York: Appleton-Century-Crofts.

Vigil, J. M., Rowell, L. N., Alcock, J., & Maestes, R. (2014). Laboratory personnel gender and cold pressor apparatus affect subjective pain reports. *Pain Research and Management, 19*, e13-e18.

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*, 13-16.

Wilson, K. G., & Luciano, C. (2002). *Terapia de aceptación y compromiso. Un tratamiento conductual orientado a los valores* [Acceptance and commitment therapy. A behavioral treatment oriented to values]. Madrid: Pirámide.

Table 1

Comparison of the Methodological Characteristics of the Studies Conducted by Sierra et al. (2016) and Pendrous et al. (2020)

	Sierra et al. (2016)	Pendrous et al. (2020)
Random allocation of the participants to the conditions	YES (p. 269)	YES (p. 20)
Experimenters blinded	YES (p. 272)	YES (p. 20)
Stratified by sex	YES (p. 269)	YES (p. 20)
Same percentage of men and women	YES (p. 269)	NO[1] (p. 21)
Participants' knowledge of ACT/RFT	NO (p. 268)[2]	YES (p. 21)[3]
Previous experience with the cold pressor task	NO (p. 268)	YES (p. 21)[4]
Heterogeneity of participants and compensation for participation	NO (p. 269)	YES (p. 19)
Percentage of participants who were exposed to the experimental procedure in their first language	100%[5]	78.6% (p. 21)
Temperature of the cold pressor task	4.5 to 5.5 °C	3.0 °C
Apparatus used in the cold pressor task	Artisanal apparatus	Jeio Tech Refrigerating Bath Circulator (RW-2025G)
Circulating water	YES	YES
Assessment of analogical reasoning skills	NO	YES (p. 19)
Administration of the experimental procedure	With an experimenter	Automatically through PsychoPy (p. 19)
Specific wording for manipulating augmentals	… there is the most important thing for you, this thing you dream about, the one that excites you the most and makes you vibrate…	… there is the most important thing to you - something that really excites you, **or something that you often think about**…
Specific wording for manipulating common physical properties	Very cold vs. thick, filthy, and smells like a sewer	Very cold vs. disgusting– it smells awful

Time for relational elaboration during the protocols	YES (pp. 270-272)	NO[6]
Assessment of protocol understanding	YES (p. 272)	NO (p. 23)

[1]77.5% of female participants. [2]Although it was not reported in the article, the study avoided participants with knowledge of ACT/RFT by not recruiting participants above the sixth semester, which was the moment in which students have some introduction to ACT/RFT, and by asking participants about their knowledge. [3]40.5% of participants reported having at least some knowledge of ACT/RFT. There were statistically significant differences in ACT/RFT knowledge across experimental conditions according to the analysis computed with the dataset provided in the article. [4]23.8% of participants reported having previous experience with the cold pressor task. [5]This was not reported in the study because of the almost null variability in the first language among Colombian undergraduates. [6]Although the study reports the same time intervals as in Sierra et al. (2016), the audio files shared by the experimenters show that there was no time for relational elaboration in Pendrous et al. (2020).

Table 2

Distribution of Participants with ACT/RFT Knowledge Across Experimental Conditions in Pendrous et al. (2020)

Experimental Condition	Knowledge of ACT/RFT*			
	Not at all	Somewhat	A lot	Total
Condition A	10	6	5	21
Condition B	8	10	2	20
Condition C	20	1	0	21
Condition D	12	5	4	21

*There were statistically significant differences across conditions according to the chi-squared test ($X^2(6) = 19.46$, $p = .003$).

Table 3

Distribution of Participants with Previous Experience with the Cold Pressor Task Across Experimental Conditions in Pendrous et al. (2020)

Experimental Condition	Previous experience with the cold pressor task*		
	YES	NO	Total
Condition A	6	15	21
Condition B	7	13	20
Condition C	1	20	21
Condition D	5	16	21

*There were marginally statistically significant differences according to the chi-squared test ($X^2(3) = 5.96$, $p = .11$).

Table 4

Number of Naïve Participants in Terms of ACT/RFT Knowledge and Previous Experience with the Cold Pressor Task in Pendrous et al. (2020)

Experimental Conditions	Naïve participants*		
	YES	NO	Total
Condition A	7	14	21
Condition B	4	16	20
Condition C	19	2	21
Condition D	11	10	21

*There were statistically significant differences across conditions according to the chi-squared test ($X^2(3) = 23.33$, $p < .001$).

Table 5

Duration of the Pauses in the Experimental Protocols in the Original and Replication Study

			Actual pauses in the audio files shared by Pendrous et al. (2020)			
	Pauses in Sierra et al. (2016)*	Pauses reported in Pendrous et al. (2020)*	Condition A	Condition B	Condition C	Condition D
Pause 1: Remembering cold pressor task	15 s	15 s	7 s	15 s	10 s	12 s
Pause 2: Imagine what is on the other side of the swamp	30 s	30 s	6 s	6 s	6 s	6 s
Pause 3: Imagine what would you feel when approaching the other side of the swamp	15 s	15 s	4 s	4 s	4 s	4 s
Pause 4: Would you jump and cross the swamp?	10 s	10 s	8 s	8 s	8 s	10 s

*The pauses were the same for all experimental conditions.