Psychonomic Bulletin & Review (2019) 26:647–653
https://doi.org/10.3758/s13423-019-01585-x

BRIEF REPORT

Illusory inconsistencies in judgment: Stimulus-evoked reference sets and between-subjects designs

Lim M. Leong[1] · Craig R. M. McKenzie[1,2] · Shlomi Sher[3] · Johannes Müller-Trede[4]

Published online: 19 March 2019

Abstract
Asked to judge the subjective size of numbers in a between-subjects design, participants rated 9 as larger than 221 (Birnbaum, 1999). The 9 > 221 effect seems to indicate that different stimuli evoke different contexts for comparison, and sounds a warning for the interpretation of between-subjects comparisons. We show that, contrary to appearances, the effect is not a result of stimulus-evoked reference sets. Instead, it is an artifact of the original 1–10 response scale and task instructions, which encourage a conflation of the response scale and the reference set. When ratings are expressed on a 1–1000 scale, or on a non-numerical slider scale, the effect reverses. However, we also show that stimuli *can* evoke their own comparative contexts, generating illusions of inconsistency in between-subjects designs. We report two novel findings – a 9 > 009 effect and a -2 > 2 effect – which are best explained by stimulus-evoked reference sets. Thus, while revealing that the 9 > 221 effect is an artifact of the original response scale, our study ultimately affirms Birnbaum's warning about the comparison of between-subjects ratings.

Keywords Context effect · Evoked reference set · Between-subjects design · Replication

Introduction

In an experiment reported by Birnbaum (1999), participants were asked to rate how large a single number was "on a scale from 1 to 10, where 1 = very very small" and "10 = very very large." Participants in one between-subjects condition rated the number 9, while those in the other condition rated the number 221. Comparing mean responses in the two conditions, a "less-is-more" pattern was observed: 9 was rated as larger than 221, even though presumably no participant in either condition actually believed that 9 is larger than 221.

This simple finding seems to have important implications, both psychologically and methodologically. Psychologically, Birnbaum interpreted the result as an effect of stimulus-evoked context: To complete the task, participants must select, perhaps implicitly, a comparative context, or reference set, that defines what will count as a "small" or "large" number. The number 9, Birnbaum suggested, tends to evoke a context of smaller (e.g., single-digit) numbers, relative to which 9 is large, while the number 221 evokes a context of larger (e.g., triple-digit) numbers, relative to which 221 is small. A sizable literature has demonstrated robust effects of context on subjective judgments (e.g., Parducci, 1965, 1995), and theoretical accounts (e.g., Kahneman & Miller, 1986) have posited that stimuli can recruit their own contexts for comparative evaluation. The 9 > 221 effect appears to illustrate the effect of stimulus-evoked context in a particularly dramatic way.

Methodologically, the 9 > 221 effect seems to offer a cautionary tale for the interpretation of judgments in between-subjects designs. Psychologists are often interested in the ordering of subjective ratings of different stimuli provided by different participants. But if different target stimuli, presented in isolation, evoke different comparative contexts, stimulus and context will be systematically confounded, and the ordering of between-subjects ratings may be meaningless.

Electronic supplementary material The online version of this article (https://doi.org/10.3758/s13423-019-01585-x) contains supplementary material, which is available to authorized users.

✉ Lim M. Leong
lmleong@ucsd.edu

[1] Department of Psychology, University of California, San Diego, La Jolla, CA, USA

[2] Rady School of Management, University of California, San Diego, La Jolla, CA, USA

[3] Pomona College, Claremont, CA, USA

[4] IESE Business School, Universidad de Navarra, Barcelona, Spain

But is the 9 > 221 effect really the result of different stimuli evoking different reference sets? There is an alternative explanation of the effect that does not involve stimulus-evoked contexts at all. Note that the original task has a peculiar structure, in that participants must use numbers to rate numbers. This may lead some participants to mistake the response scale (ratings from 1 to 10) for the intended reference set (numbers from 1 to 10). Indeed, response scale confusion may have been encouraged by the instructions in the original task, shown in Fig. 1a. Participants may have interpreted the clause "where 1 = very very small" to mean "where the *number* 1 counts as very, very small" (rather than as "where a *rating* of 1 corresponds to very, very small"). Similarly, they might have interpreted "where 10 = very very large" to refer to the *number* 10 rather than a *rating* of 10. Moreover, note that this confusion could only affect responses in the 9 condition, because the 1–10 response scale can be mistaken for a comparative context for 9 but not for 221. Ratings for 9, but not for 221, would then be inflated even if the stimuli 9 and 221, encountered in isolation, do not otherwise evoke different reference sets.

An initial indication that the 9 > 221 effect may be an artifact of the response scale was reported by McKelvie (2001; see also McKelvie, 2012; McKelvie, Juillet, & Longtin, 2013), who noted that the scale could have been used as a reference set. He reproduced the 9 > 221 effect in a direct replication of the original task, in which "largeness" was rated on a 1–10 scale with the original instructions, but did not find a significant difference between 9 and 221 when ratings were marked on a continuous line with no numerical labels. Nonetheless, the 9 > 221 effect – which remains much better known than the partial failure to replicate – is still frequently invoked in drawing broad lessons about reference set confounds in between-subjects designs.

In this article, we provide further evidence that, contrary to appearances, the 9 > 221 effect is not a result of stimulus-evoked reference sets. Instead, the effect is an artifact of the specific response scale that was used, and it does not replicate

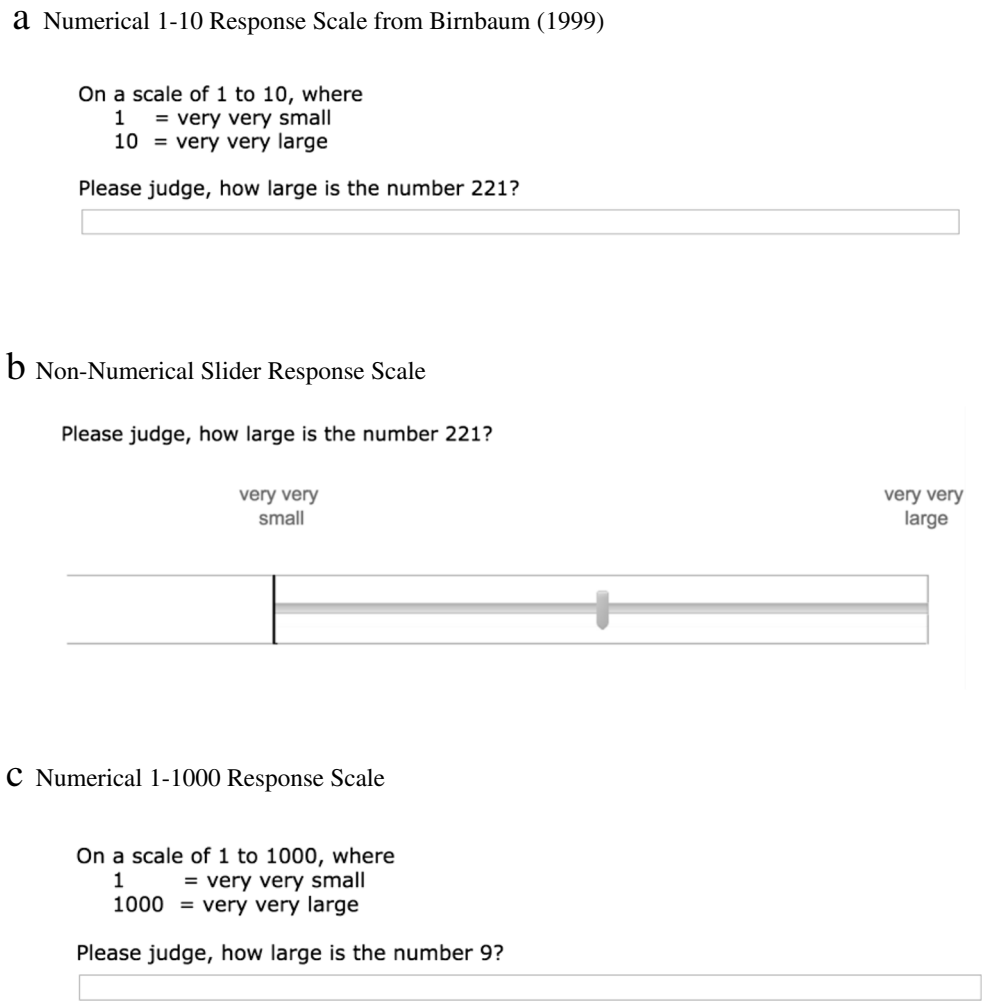

Fig. 1 The response scales used in Experiments 1–3. The original numerical 1–10 response scale (**a**) was used in Experiments 1 and 2, the non-numerical slider response scale (**b**) was used in Experiments 1 and 3, and the numerical 1–1000 response scale (**c**) was used in Experiment 2

when small changes to the elicitation method are introduced. Experiment 1 compares ratings for 9 and 221 elicited either via the 1–10 numerical scale and task instructions used by Birnbaum (1999) or via a continuous scale without numerical labels. Like McKelvie (2001), we replicate the 9 > 221 effect with the original scale and instructions. But on the continuous scale, we find that the effect does not merely disappear, as McKelvie reported, but rather reverses. Experiment 2 provides a more direct demonstration of the effect of response scale: The 9 > 221 effect again replicates using the 1–10 rating scale, but reverses when a 1–1000 scale is used. However, we also report novel findings that support the idea that individual stimuli *can* evoke markedly different evaluative contexts even when no context is suggested by the response scale, creating illusions of inconsistency in comparisons of between-subjects judgments. Experiment 3 reports two findings – a "9 > 009 effect" and a "-2 > 2 effect" – in settings that allow us to rule out an influence from the response scale or its description. Thus, while the 9 > 221 effect appears to be a product of an unusually confusing response scale, our findings ultimately underscore Birnbaum's (1999) important general warning about the pitfalls of between-subjects comparisons.

Experiment 1

Experiment 1 included both a direct replication of Birnbaum (1999) and a conceptual replication with a seemingly trivial modification of the elicitation procedure: Instead of a 1-10 response scale, participants rated numerical magnitude on a non-numerical sliding scale. McKelvie's (2001) findings suggest that the 9 > 221 effect may not generalize to the sliding scale, while a theoretical account based on stimulus-evoked reference sets entails that the effect should be robust to the elicitation method.

Method

Four hundred and fifty Amazon Mechanical Turk workers participated in exchange for $0.10, and were randomly assigned to one of four conditions. After excluding six participants who did not provide a usable response or used the same IP address as an earlier participant, we were left with a final sample of 444 participants (49.3% female, five did not report gender). Following Simonsohn's (2015) recommendation for replications, we selected a sample size per condition that was 2.5 times as large as that in Birnbaum's (1999) original study. (It is worth noting that McKelvie (2001) used a smaller sample size than Birnbaum (1999)). Our sample size of 444 results in at least 80% statistical power to detect true effect sizes of $\eta_p^2 = .017$ or larger, using a two-way ANOVA with alpha of 5%, meaning we have sufficient power to detect a small to medium or larger true effect size.

We manipulated the number that was judged (9 vs. 221) and the scale type (numerical 1–10 scale vs. non-numerical slider scale; Fig. 1). The numerical scale condition (Fig. 1a) is a direct replication of Birnbaum (1999), with the original task instructions. Our materials closely matched Birnbaum's survey, including its gender and education demographic questions. The non-numerical scale condition (Fig. 1b) employed the same question and endpoint labels, but participants indicated their response by dragging a slider along a horizontal line. The initial slider position was at the scale midpoint, and participants had to at least click the slider to make a response. No numerical values were displayed to participants, but the slider position was recorded from 1 ("very very small") to 10 ("very very large") up to two decimals for the analyses reported below. The materials and data for this experiment can be found at https://osf.io/tj478/.

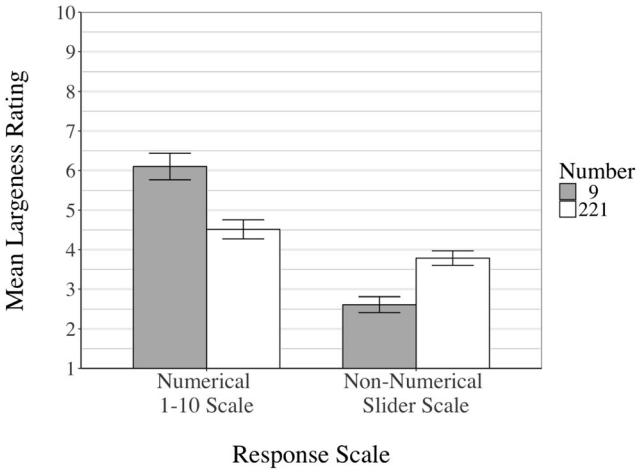

Fig. 2 Mean ratings and their standard errors for the four different conditions in Experiment 1

Results and discussion

Figure 2 shows the mean ratings for each condition. A two-way ANOVA revealed a main effect of scale type, $F(1, 440) = 73.92$, $p < .001$, $\eta_p^2 = .144$, no main effect of number, $F(1, 440) = 0.70$, $p = .40$, $\eta_p^2 = .002$, and a significant interaction, $F(1, 440) = 31.70$, $p < .001$, $\eta_p^2 = .067$. While 9 was rated as larger than 221 using the original 1–10 scale, Ms = 6.10 versus 4.51, Welch's $t(197.16) = 3.85$, $p < .001$, $d = 0.52$, this relation reversed using the non-numerical slider scale, Ms = 2.61 versus 3.79, $t(223.53) = 4.33$, $p < .001$, $d = 0.57$, with both comparisons yielding medium-sized effects (Cohen, 1988).[1]

The 9 > 221 effect depends on the response scale used. Nine is rated as larger than 221 when numerical magnitude is rated on a numerical 1–10 scale, but not when rated on a non-numerical slider scale. We, like McKelvie (2001), failed to replicate the 9 > 221 effect on the non-numerical scale – but unlike McKelvie, we obtained a significant effect in the reverse direction: 221 is,

[1] For all experiments, we report the results of Welch's t-test for comparisons between group means due to unequal variances between groups (Levene's tests, ps < .001).

reasonably enough, rated as larger than 9. The results suggest that participants rating 9 in the numerical scale condition may have simply mapped 9 to 9 – a mapping that would be unavailable in ratings of 221 on the 1–10 scale and in ratings of both numbers on the non-numerical scale. An analysis of individual responses corroborates this explanation: 54.5% (60/110) of participants gave 9 a rating of 9 in the 1–10 scale condition, while only 3.5% (4/113) rated 9 in the corresponding segment of the non-numerical scale (i.e., between 8.50 and 9.49 in recorded values), $\chi^2(1, N = 223) = 70.87$, $p < .001$, $\varphi = .56$.

Birnbaum's (1999) account of stimulus-induced contexts would predict a 9 > 221 effect with both response scales, but we found that the effect reverses when a non-numerical response scale is used. Though we cannot rule out the possibility that the two stimuli evoked different contexts for some participants, any effect of stimulus on evoked reference sets was evidently small relative to the absolute difference between the two numbers.[2] The 9 > 221 effect seems to be specific to the original elicitation procedure, in which one numerical stimulus (9), but not the other (221), has a corresponding numerical rating on the 1–10 scale. Experiment 2 supplies a more direct test of this hypothesis.

Experiment 2

In Birnbaum's (1999) paradigm, some participants appear to conflate the numerical 1–10 response scale with a comparative context for judgment when rating the number 9 but not 221. This suggests that the 9 > 221 effect should disappear when the response scale is a possible reference set for both stimuli. In particular, a 1–10 rating scale should lead to a 9 > 221 effect while a 1–1000 rating scale should lead to a 221 > 9 effect in between-subjects ratings. Experiment 2 tests this prediction.

Method

Four hundred Amazon Mechanical Turk workers participated in exchange for $0.10, and were randomly assigned to one of four conditions. After excluding four participants with duplicate IP addresses, we were left with a final sample of 396 participants (49.2% female, eight did not report gender). We manipulated the number that was judged (9 vs. 221) and the endpoints of the response scale (1–10 vs. 1–1000). Again, the two 1–10 conditions (Fig. 1a) are a direct replication of Birnbaum (1999). The 1–1000 conditions (Fig. 1c) used the same materials and instructions; our only modification was to change the response scale's upper endpoint to 1000. The materials and data for this experiment can be found at https://osf.io/tsmej/.

[2] Other features of the task context (such as the "very very small" and "very very large" endpoint labels) may affect the reference set that is evoked, but these do not differ across the two number conditions.

Results and discussion

Figure 3 shows the mean ratings for each condition, with responses on the 1–1000 scale divided by 100 to make the ratings comparable. A two-way ANOVA revealed a main effect of response scale, $F(1, 392) = 281.78$, $p < .001$, $\eta_p^2 = .418$, no main effect of number, $F(1, 392) = 0.12$, $p = .73$, $\eta_p^2 = .0003$, and a significant interaction, $F(1, 392) = 59.35$, $p < .001$, $\eta_p^2 = .131$. As in Experiment 1, 9 was rated as larger than 221 on the original 1–10 response scale, Ms = 6.03 versus 4.31, $t(181.66) = 4.10$, $p < .001$, a medium-sized effect, $d = 0.58$. By contrast, 221 was rated as larger than 9 when a 1–1,000 response scale was used, Ms = 2.18 versus 0.29, $t(192.17) = 9.74$, $p < .001$, a large effect, $d = 1.39$.

In short, 9 is only judged greater than 221 on a numerical 1–10 response scale. An analysis of individual responses provides further support for the scale-artifact hypothesis, with participants frequently providing the rating that exactly matched the number they were evaluating: 53/102 (52.0%) rated 9 a "9" on the 1–10 scale, 56/98 (57.1%) rated 9 a "9" on the 1–1000 scale, and 52/98 (53.1%) rated 221 a "221" on the 1–1000 scale. Many participants seem to interpret the numerical response scale as the reference set for evaluating the numbers. These findings provide strong evidence that idiosyncratic features of the task are driving the 9 > 221 effect, and that the effect does not simply reflect distinct stimulus-evoked contexts.

In Experiment S1 reported in the Supplementary Material available online, we dissect several further variants of the 1–10 ratings task, which differ in the wording of the instructions and the format of the response. In all variants, a substantial fraction of participants map 9 to 9 on the 1–10 scale, but only when the original instructions are used is this tendency strong enough to generate a 9 > 221 effect in mean ratings, and the effect is most pronounced when, as in the original experiment, participants must freely generate a response. The conditions that underlie the emergence of this effect are thus highly

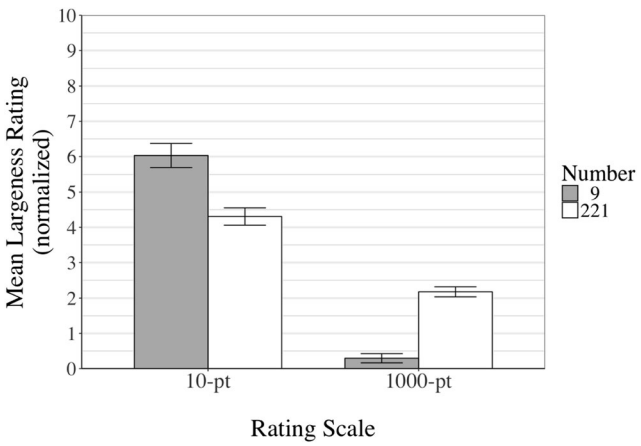

Fig. 3 Mean ratings and their standard errors for the four different conditions in Experiment 2

specific, and unlikely to generalize to typical between-subjects experiments, in which the rating scale and reference set are not confusable.

Nonetheless, the proposal that stimuli *can* recruit reference sets for evaluation has deep roots in the psychological literature (Birnbaum, 1982; Kahneman & Miller, 1986; Stevens, 1958). It should therefore be possible to leverage the recruitment of different reference sets to yield apparently inconsistent judgments in between-subjects designs. In short, while it is not exemplified by the 9 > 221 finding, we believe that the general contention advanced by Birnbaum (1999) – that isolated stimuli can evoke different reference sets, leading to apparent inconsistencies in between-subjects judgments – is likely to be valid and important. In Experiment 3, we provide support for this contention, in paradigms where it is not possible to interpret the response scale as the reference set.

Experiment 3

In Experiment 3, we employed two new stimulus sets that we expected to more strongly evoke distinctive contexts for comparison. To rule out numerical response scale effects, all conditions in Experiment 3 featured non-numerical slider scales. In Experiment 3A, we compared magnitude ratings for 9, 221, and 009. We expected a comparison of the 9 and 221 conditions to replicate the non-numerical slider condition in Experiment 1, with higher ratings for 221. The 009 stimulus, however, carries a powerful pragmatic suggestion of a three-digit reference set. This should lead to lower ratings for 009 than 9, which we expect to be less likely to evoke such a large reference set. In Experiment 3B, participants rated one of two stimuli: -2 or 2. Evaluations of -2 necessarily require a comparative context that includes negative numbers, while evaluations of 2 do not. This down-shifted context should, in turn, elevate evaluations of -2, leading to higher ratings for the smaller number in this design.

Method

Three hundred and fifty Amazon Mechanical Turk workers participated in Experiment 3A, and a different group of 550 workers took part in Experiment 3B, in exchange for $0.10. In Experiment 3A, we were left with a final sample of 346 participants (38.2% female, three did not report gender) after excluding four responses that were unusable or from duplicate IP addresses, and in Experiment 3B, we were left with a final sample of 527 participants (47.6% female, two did not report gender) after excluding 23 responses using the same exclusion criteria. The sample size in Experiment 3A was again determined using Simonsohn's (2015) 2.5 heuristic. Our sample size of 346 results in at least 80% statistical power to detect true effect sizes of η^2 = .027 or larger, using a one-way ANOVA with alpha of 5%, meaning we have sufficient power to detect a small to medium or larger true effect size. In Experiment 3B, the sample size was calculated to achieve 95% power based on the effect size obtained in Experiment S2, reported in the Supplementary Material available online.

All participants saw a single number and were asked to judge its magnitude on a sliding scale, using the method depicted in Fig. 1b. Experiment 3A had three between-subjects conditions, in which participants evaluated the number 009, 9, or 221. Experiment 3B had two conditions, in which participants evaluated the number -2 or 2. The materials and data for these two experiments can be found at https://osf.io/ym3bj/ (Experiment 3A) and https://osf.io/puazk/ (Experiment 3B).

Results and discussion

Figure 4 shows the mean ratings for each condition in Experiment 3A. A one-way ANOVA indicated a significant difference between the numbers, $F(2, 343) = 38.62$, $p < .001$, $\eta^2 = .184$. Two-sample t-tests showed that while the mean rating for 9 was smaller than that for 221, Ms = 2.43 versus 3.93, $t(222.28) = 5.83$, $p < .001$, $d = 0.77$, the mean rating for 9 was larger than that for 009, Ms = 2.43 versus 1.94, $t(216.00) = 2.27$, $p = .024$, though with a smaller effect size, $d = 0.30$. Thus, replicating Experiment 1, 221 is rated as greater than 9 on a non-numerical slider scale, suggesting that 9 and 221 fail to evoke sufficiently different reference sets to overcome the absolute difference between the numbers. Nine, however, is rated above 009, suggesting that the explicit marking of leading zeros in the latter stimulus does evoke a reference set of larger numbers.

Figure 5 shows the mean ratings by condition in Experiment 3B. Mean ratings of -2 significantly exceeded ratings of 2, Ms = 2.37 versus 1.58, $t(438.82) = 7.76$, $p < .001$, a medium-sized effect, $d = 0.68$. Whereas 2 may have evoked a reference set of only positive numbers in at

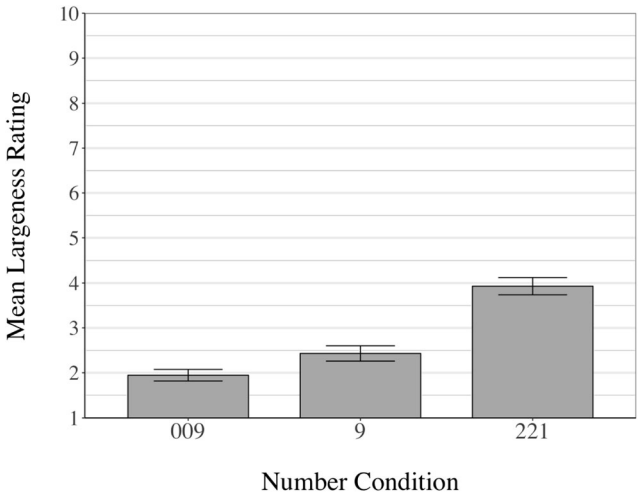

Fig. 4 Mean ratings and their standard errors for the three number conditions in Experiment 3A

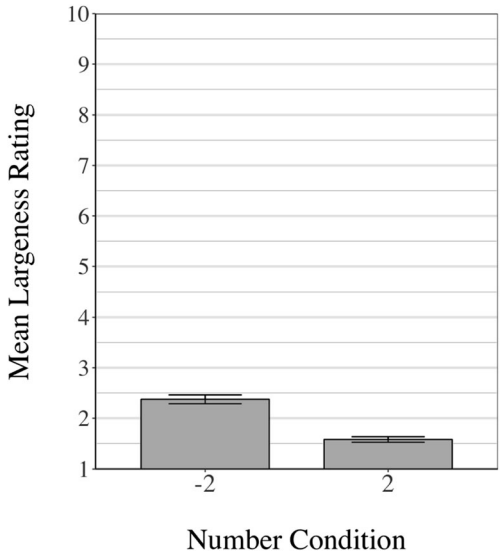

Fig. 5 Mean ratings and their standard errors for the two number conditions in Experiment 3B

least some participants, the expanded reference set of positive and negative numbers presupposed in ratings of -2 presumably led to larger ratings for the smaller number.

When the response scale cannot be interpreted as defining the reference set, 9 is not rated as larger than 221. We nonetheless find both a 9 > 009 effect and a -2 > 2 effect. These effects strongly suggest that different stimuli can indeed evoke different reference sets for comparison, leading to divergent applications of a subjective rating scale, and hence to seemingly inconsistent judgments. The results of Experiment 3 thus support Birnbaum's (1999) broader account of stimulus-evoked reference sets, as well as his broader warning about the pitfalls of between-subjects comparisons. While between-subjects size ratings of -2 exceed ratings of 2, and those of 9 exceed those of 009, it is safe to assume that all participants actually understand that 2 is larger than -2 and that 9 and 009 are equal in size.

General discussion

The experiments reported here show that the 9 > 221 effect is an artifact of the original response scale, and does not reflect the comparative contexts that stimuli spontaneously evoke. When the 1–10 response scale, which can be conflated with the comparative context for 9 only, is replaced with a non-numerical slider scale or a 1–1000 scale, the relation reverses. The detailed wording of the task instructions in the original paradigm likely encouraged such a conflation of the response scale with the comparative context. Nonetheless, we also show that numerical stimuli *can* invoke distinctive reference sets for comparison: 9 is rated as larger than 009, and -2 is rated as larger than 2, in between-subjects designs where the non-numerical response scale cannot be conflated with the comparative context.

The numerical stimuli 009 and -2 were chosen because they provide strong pragmatic cues (leadings zeros) or reference set constraints (the inclusion of negative numbers) that we believed would generate apparent inconsistencies in between-subjects comparisons. Naturally, we would expect the effect to hold for other numbers with these cues (e.g., 007, -5) as well. The findings provide a clear existence proof that different target stimuli can evoke different reference sets when other features of the task are held constant. But the present study also illustrates that it is not always *a priori* obvious when stimuli will evoke reference sets sufficiently different as to generate between-subjects ratings reversals: This initially seemed plausible for the numbers 9 and 221, but turned out not to be the case. Ultimately, the reference sets that specific stimuli evoke must be identified via empirical investigation. We have no theoretical reason to believe that the present findings depend on special features of the population, provided that participants are familiar with the relevant cues.

Stimulus-evoked reference sets may also shed light on a number of otherwise puzzling findings in the literature. For example, Slovic, Finucane, Peters, and MacGregor (2002) showed that adding a possible small loss to a gamble that offers a chance of a moderate gain can increase its rated attractiveness, an apparent violation of the normative principle of dominance. While Slovic and colleagues offered an affect-based explanation of this effect, McKenzie and Sher (2018) report evidence that suggests the effect instead results from the distinct comparative contexts that the two gambles evoke (gambles involving only wins vs. gambles involving both wins and losses). Similarly, McGraw, Larsen, Kahneman, and Schkade (2010) argued that failures to find evidence for loss aversion in between-subjects ratings of gambles involving either gains or losses reflect the masking effect of incompatible stimulus-evoked contexts: The greater impact of losses over gains is obscured in ratings because potential losses are implicitly evaluated in comparison to other potential losses only. In these examples, the stimulus appears to evoke a context of similar stimuli for the purpose of comparison.

Just as a single stimulus can evoke different comparative contexts, a small sample of stimuli may likewise trigger inferences about the real-world distribution from which it was drawn, resulting in different beliefs, and hence different evaluations, when different samples are encountered in a between-subjects design. Such sample-based inferences can create the illusion of "preference reversals" in joint-separate experiments (Hsee, Loewenstein, Blount, & Bazerman, 1999) without any true change in the ordering of options (Sher & McKenzie, 2014). Different inferences from different samples can also induce rational "intransitive choice cycles," even if underlying preferences are always transitive (Müller-Trede, Sher, & McKenzie, 2015). The different contexts that different stimuli recruit, and the different contextual inferences they support, can thus both generate illusory inconsistencies and obscure real distinctions in studies of between-subjects evaluations.

Our findings illustrate the importance of both direct and conceptual replications, as well as the thorny challenges that can arise in their interpretation (Nosek & Errington, 2017; Zwaan, Etz, Lucas, & Donnellan, 2018). While our direct replications successfully reproduced the 9 > 221 effect, this success turns out to mean little on its own: Seemingly trivial modifications to the response scale suffice to upend the effect and invalidate its original theoretical interpretation. These "unsuccessful" conceptual replications contribute to our understanding of the effect by highlighting a critical factor (the response scale) whose role was initially overlooked. But that is not the end of the story. Subsequent conceptual replications that employ stronger manipulations reaffirm the main theoretical hypothesis of the original study: Different target stimuli can evoke different evaluative contexts, leading to ordinal reversals of judgment in between-subjects designs. Together, the successful and unsuccessful replications reported here map out the remarkable complexity that lies behind an apparently simple judgment, on an apparently simple scale, of an apparently simple stimulus in apparent isolation.

Author Note This research was supported by a Scholar Award from the James S. McDonnell Foundation to Shlomi Sher.

Data and materials can be found at the Open Science Framework page for this article: https://osf.io/wdtv7/.

References

Birnbaum, M. H. (1982). Controversies in psychological measurement. In B. Wegener (Ed.), *Social attitudes and psychophysical measurement* (pp. 401-485). Hillsdale, NJ: Erlbaum.

Birnbaum, M. H. (1999). How to show that 9 > 221: Collect judgments in a between-subjects design. *Psychological Methods, 4*(3), 243-249.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, 2nd edn. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hsee, C. K., Loewenstein, G. F., Blount, S., & Bazerman, M. H. (1999). Preference reversals between joint and separate evaluations of options: A review and theoretical analysis. *Psychological Bulletin, 125*(5), 576-590.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review, 93*, 136-153.

McGraw, A. P., Larsen, J. T., Kahneman, D., & Schkade, D. (2010). Comparing gains and losses. *Psychological Science, 21*(10), 1438-1445.

McKelvie, S. J. (2001). Factors affecting subjective estimates of magnitude: When is 9 > 221? *Perceptual and Motor Skills, 93*, 432-434.

McKelvie, S. J. (2012). Exploring a counterintuitive finding with methodological implications: Why is 9> 221 in a between-subjects design. *International Journal of Humanities and Social Science, 2*, 45-51.

McKelvie, S. J., Juillet, D. R., & Longtin J.-A. V. (2013). Comparing the perceived size of 9 with 221 and with 2143: Biasing effects of inferred context in a between-subject design. *Journal of Scientific Psychology, December*, 25-44.

McKenzie, C. R. M., & Sher, S. (2018). *Gamble evaluation and evoked reference sets: Why adding a small loss to a gamble increases its attractiveness*. Manuscript submitted for publication.

Müller-Trede, J., Sher, S., & McKenzie, C. R. M. (2015). Transitivity in context: A rational analysis of intransitive choice and context-sensitive preference. *Decision, 2*(4), 280-305.

Nosek, B. A., & Errington, T. M. (2017). Making sense of replications. *eLife 6*: e23383.

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review, 72*(6), 14-418.

Parducci, A. (1995). *Happiness, pleasure, and judgment: The contextual theory and its applications.* Mahwah, NJ: Erlbaum.

Sher, S., & McKenzie, C. R. M. (2014). Options as information: Rational reversals of evaluation and preference. *Journal of Experimental Psychology: General, 143*, 1127-1143.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26*(5), 559-569.

Slovic, P., Finucane, M., Peters, E., & MacGregor, D. G. (2002). The affect heuristic. In T. Gilovich, D. Griffin, & D. Kahneman (eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 397-420). Cambridge: Cambridge University Press.

Stevens, S. S. (1958). Adaptation-level vs. the relativity of judgment. *American Journal of Psychology, 71*, 633-646.

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41*, e120.