Comparing the Perceived Size of 9 with 221 and with 2143: Biasing Effects of Inferred Context in a Between-Subjects Design

Stuart J. McKelvie, David R. Juillet and Jo-Anne V. Longtin
Bishop's University

Abstract

To investigate whether decision-making can be biased by context, different participants judged the size of different numbers that had been hypothesized to induce different contexts. To measure these contexts, the Juillet Measure of Inferred Context (JMIC) was proposed, validated and then employed to compare contexts invoked by the numbers. In Study 1, where 9 was rated as greater than 221 on a numerical scale but similar to 221 on a continuous line (McKelvie, 2001), JMIC scores were smaller for 9 than for 221 in both cases. However, the difference was reduced on the line compared to the numerical scale. In Study 2 (reported here), 9 was rated as similar in size to 2143 on both scales. However, from the JMIC scores, the context invoked by 9 was smaller than the context invoked by 2143 on both scales. The JMIC results show that between-subjects designs do not eliminate context effects. Moreover, in conjunction with the context size priming model of number judgment (proposed here), according to which (a) induced context is a positive function of the target number and (b) the numerical scale has a priming effect on this context, the JMIC results help to explain the anomalous findings that 9 was judged to be larger or similar to 221 (Study 1) and similar to 2143 (Study 2).

Keywords: Cognitive biases, number judgment, context effects, between-subjects designs, within-subject designs, rating scales

The general question of how numbers are processed has been addressed in studies of numerical cognition, which includes calculation, counting, subitizing, judging, and estimation (e.g., Lechelt, 1971); the psychophysics of number (e.g., Ganor-Stern, 2013; Thompson & Opfer, 2010); and the neuropsychological basis of number (e.g., Rickar, Romero, Basso, Wharton, Flitman, & Grafman, 2000; Vuokko, Niemivirta, & Hrlenius, 2013).

In particular, for judgment of frequency, three kinds of cognitive heuristic, or mental shortcuts, have been identified (Tversky & Kahneman, 1974). Such heuristics can assist in accurate judgment, but they may also lead to biased estimates (Galotti, 2004; Randell, 2009). With the *representative heuristic*, the judgment is based on how typical the event seems to be. More typical events are judged as more frequent. With the *availability heuristic*, the judgment is based on how easily the event comes to mind. Events that come more easily to mind are judged as more frequent. With the *anchoring-and-adjustment heuristic*, the judgment is adjusted from an initial value. The person begins with a number (the anchor), which then exerts a systematic influence on the subsequent judgments, pulling them closer to the anchor and thereby increasing or decreasing frequency estimates (Epley & Gilovich, 2006). For example, according to Epley & Gilovich, when people were asked if they think that the population of Chicago is more or less than 200,000 or more or less than 5 million, subsequent estimates are lower in the first case than in the second case. The lower anchor (200,000) pulls the estimate down and the higher anchor (5 million) pulls the estimate up. The anchoring effect can also occur when numbers themselves are judged. Estimates for the product of the numbers 1 to 8 were lower when shown as 1 X 2 X 3 etc. than when they were shown as 8 X 7 X 6 etc. (Tversky & Kahneman, 1974). The initial numbers served as an anchor, pulling the estimated products down or up, respectively.

The main reason why anchoring causes bias is that the initial value provides a *context* for the subsequent judgments. Consequently, the same target may be judged differently depending on the context in which it occurs.

Correspondence concerning this article should be addressed to Stuart McKelvie, PhD, Department of Psychology, Bishop's University, 2600 rue College, Sherbrooke (Borough of Lennoxville), QC, J1M 1Z7, Canada.; email: smckelvi@ubishops.ca.

The effect of context was vividly demonstrated by Birnbaum (1974), who asked people to judge the size of 45 numbers on a verbally-labeled scale with 9 steps from very very small to very very large. Estimates were generally higher when they were presented in a positively-skewed distribution for some people than when they were presented in a negatively-skewed distribution for others. Of particular interest, when 450 appeared in the positively-skewed distribution it was rated higher than 550 in the negatively-skewed distribution. This result clearly violates the natural ordering of numbers. Notably, it has been argued that the psychophysical relationship relating subjective number to objective number is captured by Fechner's Law (Thompson & Opfer, 2010) and Weber's Law (Whalen, Gallistel, & Gelman, 1999). This means that subjective number is a negatively accelerated function of objective number (see also McKelvie & Shepley, 1977). That is, the psychological distance between numbers is greater with smaller numbers than with larger numbers. However, even if the subjective distance between 450 and 550 is smaller than, say, between 100 and 200, the finding that 450 is judged to be greater than 550 is clearly anomalous.

Mellers and Birnbaum (1983) extended this result to social judgment when they showed that evaluations of student performance were also affected by whether the distribution of examination scores was positively skewed or negatively skewed. These findings with numbers and with social judgments suggest that there are many subtle context effects on judgment. Indeed, Mellers and Birnbaum (1982) argue that all judgments are relative, and that a complete theory of psychophysical judgment must take context into account.

Between- vs. Within-Subjects Designs

When Birnbaum (1974) found that 450 was rated as greater then 550, and indeed with the other examples of anchoring and context effects that were cited above, the results were obtained with between-subjects design. That is, the judgments were made by different groups of people who rated one of the two targets in different contexts. This implies that the research design itself may be a source of bias. In fact, a perennial methodological problem in psychology is whether experiments should be conducted with participants taking part in different conditions (a between-subjects design) or in all conditions (a within-subjects design) (Greenwald, 1976). Writers usually identify the main difficulty with between-subjects designs as the subject selection effect (Christensen, Johnson, & Turner, 2011; McBurney & White, 2004). Although researchers attempt to control subject error variance by randomization, matching, or both, there is no guarantee that the error is eliminated, meaning that treatment effects may be confounded in unknown ways by individual differences. This problem can be eliminated by the within-subjects design because the same people serve in all conditions, which provides a more sensitive test of the effect of the independent variable. In addition, this design is highly efficient because fewer participants may be required.

However, taking part in more than one condition can be taxing, can lead to new difficulties such as order and carry-over effects (Christensen, et al., 2011; Keren & Raaijmakers, 1988; Poulton, 1973), and may even permit participants to discover the researcher's hypothesis (Keren & Raaijmakers, 1988). Although doubt has been expressed about the latter claim that within-subjects designs are transparent (Lambdin & Shaffer, 2009), and techniques such as counterbalancing can tackle many of the sequence effects (Christensen, et al., 2011; McBurney & White, 2004; Poulton, 1973), these techniques may not always be successful (Keren & Raaijmakers, 1988). For example, counterbalancing may not remove transfer effects if they are asymmetrical (Poulton, 1973), which means that the second judgment will be biased by the first one. Together, these considerations dictate that care must be taken when choosing a within-subject design or a between-subjects design.

In addition to the debate about the methodological advantages and disadvantages of the two designs, there is concern that they might lead to different results. For example, for the action effect conundrum, in which regret over a negative outcome is greater after performing an action than after deciding not to act, Zhang, Walsh and Bonnefon (2005) found that the effect was much weaker with a between-subjects design than with a within-subjects design. In addition, for the framing effect, in which judgments of risk are a function of whether outcomes are phrased in positive or negative terms, the effect has been stronger with between-subjects designs than with within-subjects designs (Mahoney, Buboltz, Levin, Doverspike, & Svyantek, 2011).

When participants take part in more than one condition or make more than one kind of judgment, their initial experience will serve as a baseline or a context for their subsequent experiences. In some cases, for example if the task requires a direct comparison between two or more targets, this is acceptable, and the context itself may even be of interest (Greenwald, 1976). However, in other cases, the earlier experience can lead to biased results. This was seen in the anchoring effect and in the sequence effects that were discussed above. Because of problematical context effects in within-subjects designs, it has been suggested that these designs should be avoided or that results obtained with them should routinely be checked with a between-subjects design (Poulton, 1973). However, as seen with the judgments of the size of numbers in different contexts (Birnbaum, 1974), a between-subjects design does not necessarily eliminate context effects (see also Greenwald, 1976).

The Interesting Case of 9 > 221

Birnbaum's (1999) Experiment

Although 450 was rated greater than 550 in his between-subjects design (Birnbaum, 1974), Birnbaum (1999) reported a dramatic and even more counterintuitive result that further questions whether between-subjects designs eliminate context effects. Using a between-subjects design, but this time with a numerical scale from 1 to 10 in which the endpoints were labeled as 1 = *very very small* and 10 = *very very large*, participants rated the size of the number 9 or the number 221. The mean rating for 9 (5.13) was significantly higher than the mean rating for 221 (3.10). Both of these results violate the natural order of the numbers, but the second result is particularly anomalous because not only is the difference between 9 and 221 greater than the difference between 450 and 550, Fechner's Law shows that the psychological distance between numbers is greater with smaller numbers than with larger numbers.

To explain his startling finding, Birnbaum (1999) argued that each number induced its own context: 9 induced the context of single digits and 221 induced the context of triple digits. For Birnbaum, 9 was then perceived as high in its context and 221 was perceived as low in its context, leading participants to rate 9 as larger than 221. It appears that the between-subjects design itself was the source of a context effect that had a biasing effect on judgment. Birnbaum's model, which postulates a *specific numerical context for each number*, will be referred to as the *specific context model*.

McKelvie's (2001) Model and Experiment

However, McKelvie (2001) questioned Birnbaum's account that each number induced its own context and suggested that 9 might have been rated relatively high because participants interpreted the label "very very large" to *literally mean* the number 10. This implies that 9 was rated higher than 221 because it was perceived to be close to the number (10) that was by definition very very large. In addition, McKelvie suggested that the labels very very small and very very large that were chosen for the endpoints of the rating scale may have induced the context of a *wide range of numbers*, perhaps even all positive numbers, and that 221was regarded as a low number in that context. In other words, the rating of 221 was a natural judgment but the higher rating of 9 was a spurious consequence of the numerical scale. McKelvie's model, which *postulates a wide context in general and a spurious effect of the numerical scale*, will be referred to as the *wide context spurious model*.

McKelvie's (2001) experiment: Part 1. To compare these accounts, McKelvie replicated Birnbaum's experiment with his original 1 to 10 numerical scale but added a second condition in which ratings were made on an 87 mm continuous line. The endpoints were again labeled as very very small and very very large, but there were no numbers on the scale. Ratings were scored by measuring their distance from the left hand side of the line and by converting them into numbers from 1 to 10. McKelvie argued that if Birnbaum's specific context model was correct, 9 would be rated as greater than 221 on both scales because 9 would be judged as large relative to single digit numbers and 221 would be judged as small relative to three digit numbers. In contrast, if McKelvie's wide context spurious context model was correct, 9 would again be rated as higher than 221 on the numerical scale. However, this would occur because 9 was judged as close to 10 which was the very very large number and 221 was low because it was judged in the context of a wide range of numbers, not simply 3-digit numbers. Most importantly, on the continuous line, 9 and 221 would *both* be rated as relatively low because they would both be absolutely low in the context of the same wide range of numbers.

The results were clear. Replicating Birnbaum's result with his numerical scale, 9 (mean rating of 6.58) was judged to be greater than 221 (mean rating of 3.21). However, on the continuous line, 9 (3.80) and 221 (3.52) received similar fairly low ratings. McKelvie suggested that Birnbaum's higher rating of 9 compared to 221 in Birnbaum's (1999) experiment and in McKelvie's (2001) experiment was a spurious consequence of the numerical scale, where the number 10 was taken literally to mean very very large. He also suggested that, on the continuous line, 9 and 221 were both judged to be relatively small in the context of a wide range of numbers. In addition, on the numerical scale, 221 was judged as small for the same reason.

An alternative version of McKelvie's (2001) model: a precise priming effect from the numerical scale. There is an alternative account for how the ratings of 9 might have occurred on the numerical scale. The presence of 1 and 10 as numerical values for the verbal endpoints of the scale, along with the requirement to report the judgment as a rating between 1 and 10, may have initially *primed the idea of a numerical context*, after which, as Birnbaum argued, 9 induced the context of single digits and 221 induced the context of triple digits. However, on the continuous line, where there were no numbers, 9 and 221 would not induce their own contexts. Rather, as McKelvie (2001) suggests, they would both be judged as fairly small in the same wide context implied by the labels very very small and very very large. The idea that the rating scale numbers have a priming effect is consistent with other evidence that the mere presence of numbers can automatically affect cognitive processing, even though the value of

the numbers is not relevant to the main task (Dehaene & Akhavein, 1995). For example, in the size congruity effect, judgments of the physical size of the fonts in which two numbers appear are affected by the numerical distance between the numbers (Choplin & Logan, 2005). This version of McKelvie's model, which postulates *a specific priming effect of the numerical scale*, will be referred to as the *wide context specific priming model*.

Comparing the three models. It is possible to derive precise predictions for the judgments of 9 and of 221 on each scale from the three explanatory proposals (see Table 1, Model Predictions, Number Judgments). According to Birnbaum's (1999) *specific numerical context model*, which applies to both scales, 9, as the highest single digit, would receive a rating of 10. For 221, the context will be the numbers 100 to 999. In fact, 221 lies .1346 of the distance between 100 and 999 [(221 – 100)/899]. On the scale from 1 to 10, this would mean that 221 would receive a rating of 1 + (9 X .1346) = 2.21, which is close to 2.

According to McKelvie's (2001) *wide context spurious model*, on the numerical scale, 10 is taken literally to be what is meant by a very very large number, which means that 9 would be rated as 9 because it is the closest number to 10. However, when 221 is judged, 10 is unlikely to be seen literally as very very large because the presence of 221 implies that very very large must exceed 10. Here, the assumed context is a wide range of numbers, perhaps all numbers, so that 221 would receive a very low rating, probably 1. On the continuous line, the assumed context for McKelvie is the wide range of numbers for *both* 9 and 221. This would mean that both 9 and 221 would receive a rating of 1.

Finally, according to the *wide context specific priming model*, 9 and 221 induce their own digital contexts, but only on the numerical scale. On this scale, 9 would be rated as 10 and 221 would be rated as 2, just as Birnbaum predicts. However, on the continuous line, both would be rated as 1, as McKelvie predicts. The predictions for McKelvie's wide context spurious model and for the wide context specific priming model are very similar (see Table 1). The models differ mainly on the *reason* why 9 would be judged to be greater than 221 on the numerical scale.

How do the results from Birnbaum (1999) and McKelvie (2001) (see Table 1, Obtained Results, Number Judgments) compare with the (specific) predictions from the three proposals? The higher score for 9 than for 221 on the numerical scale that was obtained by both Birnbaum and McKelvie is consistent with the direction of the predictions from all three proposals, and the nonsignificant difference between 9 and 221 on the line is consistent with the lack of a difference predicted by the wide context spurious model and the wide context specific priming model. However, on the numerical scale, the differences in the two experiments between the rated size of 9 and of 221 (5.1 – 3.1 = 2, in Birnbaum, 1999; 6.6 – 3.2 = 3.1 in McKelvie, 2001) was less than predicted (10 – 2 = 8 from Birnbaum's model, 9 – 1 = 8 from McKelvie's model) and, on the line, the scores for 9 and for 221 (3.8, 3.5) were higher than predicted by McKelvie's model (both 1). If 9 and 221 had been rated as 1, even though 221 is absolutely higher than 9, the results would have been consistent with McKelvie's model that the numbers are small in the context of all numbers. However, although the similar ratings of 3.8 and 3.5 for 9 and for 221, respectively, on the line are not as counterintuitive as the finding that 9 was judged to be larger than 221 on the numerical scale, the ratings are also anomalous because, being similar, they violate the natural ordering of the numbers.

A fourth model: the context size priming model. The results on both scales can perhaps be best accommodated by postulating a fourth model that includes the current suggestion that the numerical scale has a priming effect on induced context together with a change to Birnbaum's specific context model.

First, it is proposed that, on *both the numerical scale and on the continuous line*, 9 and 221 invoke different contexts that are *positively associated with their size*, rather than Birnbaum's specific contexts of single- and triple-digit numbers for 9 and 221, respectively. In this less-precise version of Birnbaum's model, the context for 9 is wider than single digits and the context for 221 is narrower than triple digits, but the context invoked by 9 is still smaller than the context invoked by 221.

Second, it is proposed that, for the *numerical scale*, the presence of 1 and 10 with the verbal labels at the end points, along with the requirement to report the judgment as a rating from 1 to 10, *automatically primes the idea of a numerical context for the number being judged*. However, the context not only varies positively with the number (as just suggested), but also *approaches* the specific values of single digits for 9 and triple digits for 221. Consequently, the difference between the sizes of the contexts for 9 and for 221 will be greater on the numerical scale than on the line. The more that the induced ranges approach those proposed by Birnbaum, the more likely it is that 9 will be very high within its range and that 221 will be low within its range. This fourth model, *which postulates a positive association between the target number and the size of the invoked context in conjunction with a priming effect on the numerical scale*, will be referred to as the *context size priming model.*

Notably, for the continuous line, McKelvie (2001) suggested that 9 and 221 might be judged as similar in the context of a wide range of numbers, perhaps all numbers, which implies that both 9 and 221 would be rated as 1. In order to explain why 9 and 221 were judged to be similar in size on the line but were rated as greater than 1, the

context in which 9 is judged must logically have been smaller than the context in which 221 is judged. This inference is consistent with the suggestion above that the difference between the sizes of the contexts for 9 and for 221 will be smaller on the line than on the numerical scale.

McKelvie's experiment: Part 2. After making their first rating, participants rated the other number (221 after 9 or 9 after 221). In this within-subjects design, 221 was generally rated higher than 9 in both scale conditions. That is, 221 was rated greater than 9 when it was rated after 9 and 9 was rated less than 221 when it was rated after 221. Clearly, the first rating served as the context for the second one, giving rise to the logical ordering of the perceived size of the two numbers. Notably, Birnbaum predicted this outcome because his hypothesis only applies to the initial rating in the between-subjects design. Numbers would retain their logical relationship to each other in a within-subjects design (Birnbaum, 1999).

Finally, McKelvie (2001) asked participants if they had been thinking of the numerical meaning of the endpoints of the scale when they were making their first judgment. For those answering in the affirmative (63% of the sample), the estimates for very very small did not vary greatly among the conditions, and the most frequent answer was low and often close to 1. This is in line with Birnbaum's hypothesis concerning the context for 9, which should be the digits 1 to 9, with 1 as the expected lower limit, but not for 221, where the context should be 100 to 999, with 100 as the expected lower limit. For very very large, estimates were more variable, but did not change much among the conditions. In particular, with regard to Birnbaum's model, on the numerical scale, 2 people out of 12 (17%) chose 9 or 10 for very very large when 9 was judged, and 2 people out of 21 (10%) chose close to 999 for very very large when 221 was judged. On the continuous line, 2 people out of 16 (13%) chose 9 or 10 when 9 was judged and 5 people out of 25 (20%) chose close to 999 when 221 was judged. These results offer only limited support for Birnbaum's contention that 9 invokes the context of single digits and 221 invokes the context of triple digits. At best, this may have occurred in only a few cases.

In fact, the most frequent answer for very very large was that it was greater than 1000 (and usually much larger). This is more consistent with McKelvie's contention that on the continuous line both 9 and 221 were rated relative to the same wide range of numbers, perhaps all positive numbers. At the same time, McKelvie (2001) argued that there was some evidence that the two numbers induced somewhat different contexts, at least on Birnbaum's numerical scale. When responses to very very small were classified as less than 9 or more than 9, the highest number of people for more than 9 occurred with 221 on the numerical scale. In addition, when responses to very very large were classified as less than 999 or more than 999, the highest number of people for less than 999 occurred with 9 on the numerical scale. These results suggest that the numerical scale had some differential effect on what was perceived as very very small and very very large for 9 and for 221 (McKelvie, 2001). However, they do not support Birnbaum's specific context model, according to which 9 induces the context of single digits and 221 induces the context of triple digits. Rather, they are more consistent with the context size priming model, in which smaller numbers induce smaller contexts than larger numbers.

In the earlier discussion of Birnbaum's (1999) results and of his account of them, it was observed that the numerical scale might have contributed to the higher ratings of 9 compared to 221 in two ways. Firstly, the number 10 might have been taken as the meaning of very very large, leading to a high rating for 9. However, 221 would have been rated as low within a wide context of numbers (McKelvie, 2001). Secondly, as suggested here, the presence of numbers may have primed the smaller invoked context for 9 than for 221, with 9 being higher in its context than 221 in its context. The reports for the meaning of very very small with 221 on the numerical scale are more consistent with the second account because most people stated that the meaning of very very small was more than 9. According to the first account, the lower bound for this context should be a very low number, probably 0 or 1, which is not the case.

Together, these analyses suggest that the most economical interpretation of the results of the experiments by Birnbaum and by McKelvie is the context size priming model proposed above: 9 invokes a smaller context than 221 on both scales; however, due to the numerical priming effect on invoked context, the difference between the size of these contexts is smaller on the line than on the numerical scale.

The models discussed here all emphasize the effect of context on number judgment in the present task. Consequently, one purpose of the present paper is to propose a measure of inferred context for this task, to employ this measure to examine if inferred context varies with the size of the numbers, and to compare Birnbaum's (1999) model, McKelvie's (2001) model, and the modified model as explanations for the anomalous number judgments.

Proposing a Measure of Inferred Context

As already noted, only 63% of McKelvie's (2001) participants reported thinking of the meaning of the endpoints. The others may not have thought about it, may not have remembered thinking about it, or may have decided not to report what they thought. However, it would be useful to have information about the contexts that all participants were working with, whether consciously or nonconsciously, when they made both of their judgments on

the scale. Any measure of these contexts would have to permit predictions from Birnbaum's model that, when presented first, 9 induces the context of single digits (a smaller context) and 221 induces the context of triple digits (a larger context), and also from McKelvie's model that both numbers were simply judged in the context of a wide range of numbers (a large context).

Here, we propose that such a measure can be derived from the two judgments that were made in the within-subjects part of the design in McKelvie's (2001) experiment. For the first number, it is assumed that a context is invoked and that the number is judged within that context. Consequently, as the invoked context becomes larger, the rating for the target number will be smaller, and there will also be a greater distance from the rating to the top of the scale. For example, if 5 was the target number, it would be rated relatively lower in a context from 1 to 1000 (a larger context) than in a context from 1 to 9 (a smaller context), and the distance from the rating to the top of the scale will be greater for 1000 than for 9. For the second number, it is assumed that it is made relative to the first one, which means that it is also influenced by the original invoked context. Again, a larger context would be associated with a lower rating and a greater distance to the top of the scale.

The proposed measure of inferred context is based on the distance between the mean of the two judgments and the top of the scale. A larger context means a larger distance and a greater context score. Because rating scales may differ in how many scale points they have, the distance in the measure is then converted into a proportion of the maximum value on the scale.

Given that all ratings in the present study were either gathered on Birnbaum's (1999) 10-point numerical scale or were transformed to that scale from ratings on the continuous line, the *Juillet Measure of Inferred Context* (JMIC) was defined as follows:

$$10 - (J1 + J2)/2$$

where J1 is the rating made first and J2 is the rating made second.

This was then converted to a proportion of the total: $[10 - (J1 + J2)/2]/10$.

This gives $JMIC = 1 - (J1 + J2)/20$.

The minimum value of each judgment is 1 (the maximum distance from the top of the scale), giving

$$JMIC = 1 - (1 + 1)/20 = 1 - .10 = .90,$$

which is the maximum value of the context score.

The maximum value of each judgment is 10 (no distance from the top of the scale), giving

$$JMIC = 1 - (10 + 10)/20 = 1 - 1 = 0,$$

which is the minimum value of the context score.

To illustrate how the JMIC reflects the context invoked when the first number is judged and is then employed in the judgment of the second number, Table 2 shows JMIC values for a variety of possible invoked contexts for 9 rated first (followed by 221) and for 221 rated first (followed by 9). For example, in the first case, if the invoked context for 9 was the range of numbers 1 to 10, 9 might receive ratings of 8 or 9 and 221, being higher than 9, would receive a higher rating of 9 or 10, yielding a JMIC score of .05 to .15. If the invoked context for 9 was a range of numbers greater than 10,000, 9 would probably receive a rating of 1 and 221 might receive a rating of 2 (or even 1 if it was judged as extremely small like 9), yielding a JMIC score of .85 or .90. In the second case, when 221 is rated first, if the invoked context was 1 to 10, 221 would be rated as 10 and 9, being smaller than 221, might receive a rating of 8 or 9, yielding JMIC scores of .05 or .10. If the invoked context for 221 was greater than 10,000, 221 would probably receive a rating of 1 and 9, being smaller than 221, would also have to receive a rating of 1, yielding a JMIC score of .90. Table 2 shows that, in both cases, as invoked context increases, JMIC scores increase.

Because Birnbaum's (1999) specific context model, McKelvie's (2001) wide context spurious model, and the wide context precise numerical priming model are stated in exact terms, it is possible to derive precise predictions for JMIC scores from these models.

Predictions from Birnbaum's (1999) Specific Context Model

If judgments were made according to Birnbaum's specific context model, they would occur as follows. Firstly, and as observed earlier, if 9 induces the context of single digits (1 to 9), which is very small context, and 9 is rated first, 9 would receive the maximum rating of 10 because it is the highest single digit (see Table 1). The participant would also have to rate 221 as 10, because that is the closest rating that captures the fact that 221 is greater than 9, a relationship that would hold in a within-subjects design. As shown above, the measure of inferred context in this case would then be 0 (see bottom half of Table 1, Model Predictions, JMIC scores), which is small.

Secondly, if 221 induces the context of triple digits, which is a larger context, and 221 is rated first, then 221 would receive a rating of 2.21 (as shown above), which is close to 2 (see Table 1, top part). The participant would then rate 9 as 1, giving $JMIC = 1 - (2 + 1)/20 = .85$ (Table 1, bottom part). Given that .90 is the maximum, this value represents a large context. Also for Birnbaum, it would not matter whether the ratings were made on the numerical scale or on the continuous line. These predictions are also shown in Table 1 (bottom part), where it is

assumed that the original judgments on the 87 mm line are converted to Birnbaum's 1 to 10 numerical scale before being entered into the formula.

Predictions from McKelvie's (2001) Wide Context Spurious Model

If judgments were made according to McKelvie's wide context spurious model, they would occur as follows (see Table 1, bottom part). For Birnbaum's numerical scale, when 9 is judged first, it would be rated as 9 because if 10 is the very very large number then 9 is one step below that. The second rating for 221 would then have to be 10 because 221 is larger than 9. This yields a value for the JMIC of 1 – (9 + 10)/20 = .05, which is small. However, when 221 is judged first, the assumed context is a *wide range of numbers*, perhaps all numbers, which is a very large context, so that 221 would receive a very low rating, probably 1. If 221 was rated as 1, 9 would also be rated as 1, giving JMIC = .90, which is large.

For the continuous line, the assumed context for McKelvie is the wide range of numbers for *both* 9 and 221. When 9 is rated first, it would receive a low rating, probably 1. If that occurs, the participant might rate 221 as 2, because that captures the fact that 221 is greater than 9 while still recognizing that 221 is a relatively small number in the wide context. The measure of inferred context in this case would then be JMIC = 1 – (1 + 2)/20 = .85, which is a large context. When 221 is rated first, it would also probably be rated as 1, giving JMIC = .90 as shown above for the numerical scale. That is, under McKelvie's model, inferred context on the continuous line would be large for both 9 being rated first and for 221 being rated first. Again, these predictions are shown in Table 1 (bottom part).

Predictions from the Wide Context Specific Priming Model

As with number judgments, the predictions from this model follow those of McKelvie for the continuous line and those of Birnbaum for the numerical scale, where the numbers are hypothesized to induce the single-digit and triple-digit contexts for 9 and for 221, respectively. These predictions also appear at the bottom of Table 1.

Predictions from the Context Size Priming Model

Although it was argued above that the best model to account for the judgments of 9 and of 221 is the context size priming model, it was not possible to derive specific predictions for JMIC scores from this model because it does not yield exact numerical predictions for the judgments themselves. However, it can be predicted generally that JMIC scores would be smaller for 9 than for 221 and that this difference would be less on the line than on the numerical scale.

Summary of Predictions for JMIC Scores

For Birnbaum's specific context model, 9 induces a smaller context than 221 on both scales, with JMIC = 0 and .85, respectively. For McKelvie's wide context spurious model, on the numerical scale, JMIC scores would again be smaller for 9 than for 221 (.05, .90, respectively). However, on the continuous line, the context would be large for both 9 and for 221 (.85, .90). For the wide context specific priming model, JMIC scores would be smaller for 9 than for 221 (0, .85) on the numerical scale and large for 9 and for 221 (.85, .90) on the line. For the context size priming model, JMIC scores would be smaller for 9 than for 221, but the difference would be smaller on the line than on the numerical scale.

Study 1: New Analyses of McKelvie's (2001) Data

As observed above, McKelvie replicated Birnbaum's experiment in which participants judged the size of 9 or 221 on the numerical scale and repeated the procedure with different participants who made their judgments on a continuous line. Because both numbers were rated (9 then 221, or 221 then 9), it was possible to calculate JMIC for each participant. To compare the JMIC scores from the line with those from the numerical scale, the ratings on the line were measured in mm from the left hand side and then converted into numbers on the scale from 1 to 10 where 1 corresponded to 0 on the line and 10 corresponded to 87 on the line.

Validity of JMIC Scores

Before conducting the new analysis of McKelvie's (2001) data, the validity of the JMIC measure was examined by using the data for the meaning of very very large as a criterion. Given that the estimates for very very small were generally very close to 1 (63 out of 71 participants gave estimates of 0, 1, or 2), examining the number reported for very very large permits us to observe the size of the context experienced by these participants. For present purposes, their reports were classified as follows: 1 = 1 to 10, 2 = 11 to 100, 3 = 101 to 1000, 4 = 1001 to 10, 000 and 5 = 10,000 or more. This scale reflects the importance that Birnbaum attached to ranges of digits (e.g., single for 9, triple for 221).

Assuming that these numbers indicate the size of the context, they can be used to provide information on the validity of the JMIC. That is, if the JMIC is a valid measure of context, scores should correlate positively with the direct reports for the meaning of very very large. These results are shown in Table 3 for the 71 (out of 112) participants who reported on the meaning of very very large. It can be seen that although sample sizes were much

smaller in each condition, the corresponding correlations were positive, ranging from .473 to .702, and all but one (.553) were significant. These results show that JMIC scores were positively correlated with reports of the meaning for very very large, and provide supporting evidence for the validity of the JMIC as a measure of invoked context.

Analysis of JMIC Scores

Turning to the main analysis of JMIC scores, a 2 X 2 (Scale X Number) ANOVA yielded two significant effects: the main effect of number, $F(1, 108) = 70.44$, $p < .001$, $\eta^2 = .380$, and the interaction between scale and number, $F(1, 108) = 5.20$, $p = .025$, $\eta^2 = .028$. The main effect of scale was not significant, $F(1, 108) = 0.69$, $p = .41$, $\eta^2 = .004$. It can be seen from Table 1 (lower part) that JMIC scores were lower for 9 than for 221, and that the difference between the means was greater on the numerical scale (.36, .78) than on the continuous line (.48, .72). Post hoc comparisons showed that the effect of number was significant on both the numerical scale, $t(51) = 7.31$, $p < .001$, and on the continuous line, $t(57) = 4.47$, $p < .001$. The standardized effect sizes were $d = 1.95$ on the numerical scale and $d = 1.17$ on the line. Both effects are very large by Cohen's (1977) standard of 0.2 for small, 0.5 for medium and 0.8 for large, but the difference in JMIC scores is greater on the numerical scale than on the line.

Discussion of JMIC Scores

The fact that JMIC scores were lower for 9 than for 221 on both scales is more consistent with Birnbaum's (1999) specific context model than with McKelvie's (2001) wide context spurious model, according to which the score for 9 should only be lower on the numerical scale. On the other hand, the effect of number was smaller on the line than on the numerical scale, which is more consistent with the wide context spurious model and the wide context specific priming model than with Birnbaum's specific context model, according to which the effect should be similar on both scales. However, the fact that the difference between JMIC scores was less on the line (.48 vs. .72) than on the numerical scale (.36 vs. .78) is most consistent with the context size priming model. This supports the conclusion that was drawn concerning the judgments themselves.

Study 2: Is 9 > 2143?

Another purpose of the present paper was to provide additional evidence of the validity of the JMIC and to examine the scores with a different comparison, in this case between 9 and 2143. The number 2143 was chosen because its position in the range of 4-digit numbers is approximately the same as the position of 221 in the range of 3-digit numbers. As shown earlier, 221 is .1346 of the distance between 100 and 999, which is equivalent to a value of 2.21 (2) on the 1 to 10 scale. Similarly, 2143 is .127 of the distance between 1000 and 9999, which is equivalent to a value of 2.14 (2) on the 1 to 10 scale. If Birnbaum's (1999) specific context model is correct, 9 should be judged as larger than 2143 because 9 is high among single-digit numbers and 2143 is relatively low among 4-digit numbers. In addition, JMIC scores should be lower for 9 than for 2143. Indeed, as for Study 1, precise predictions can be formulated from Birnbaum's (1999) specific context model, from McKelvie's wide context spurious model (2001), and from the wide context specific priming model for both number judgments and for JMIC scores. These predictions are shown in Table 4.

For Birnbaum's specific context model, the predictions for the judgments of the size of 9 and of 2143 are the same as for 9 and 221. For McKelvie's wide context spurious model, the predictions for 9 are also the same as before. However, it was expected that 2143 might be rated as 2, rather than 1 as was predicted for 221. In view of the reports for the meaning of very very small and very very large in Study 1, where a wide range of numbers was reported but not typically all numbers, it was surmised that 2143 was unlikely to be rated at the lowest level on the scale. As before, predictions for the wide context specific priming model were a mix of the previous two. JMIC scores were also derived as before. The only difference was that if 2143 was initially rated as 2, 9 would be rated as 1, yielding a JMIC scores of $1 - (2 + 1)/20 = .85$ (compared to the predicted JMIC score of .90 for 221).

As for Study 1, it was not possible to derive precise predictions for the context size priming model. However, if it is correct, 9 would be rated as larger than 2143 on both scales, with a reduced difference on the line, and possibly no difference. In addition, JMIC scores would be smaller for 9 than for 2143, with a reduced difference on the line.

Study 2 was conducted in two parts. In Study 2A, 9 and 2143 were compared in a between-subjects design using Birnbaum's numerical scale, just as Birnbaum (1999) originally did with 9 and 221. In Study 2B, this comparison was replicated using Birnbaum's scale and then examined with the continuous line, just as McKelvie (2001) did with 9 and 221.

To plan sample size, standardized effect sizes were calculated for the comparisons of 9 with 221 on Birnbaum's scale. They were found to be $d = 0.71$ (Birnbaum, 1999) and $d = 1.20$ (McKelvie, 2001), which average out to $d = 0.95$. Setting alpha at .05 and power at .70 gave a required sample size of 15 (http://www.divms.uiowa.edu/~rlenth/Power/). Consequently, Study 2 was conducted with approximately 15 people in each experimental condition. This number is sufficient to be 70% certain to detect the effects predicted by

Birnbaum's hypothesis. In addition, it is commensurate with Birnbaum's (1999) observation that he obtained his significant difference in the judgments 9 and 221 after testing only 10 participants.

Participants

Participants were university students drawn from a variety of programs. In Study 2A, 36 participants were allocated randomly to two conditions in which 9 or 2143 was judged first using Birnbaum's 1 to 10 numerical scale. In Study 2B, 75 participants were allocated randomly to four conditions in which 9 or 2143 was judged first using the numerical scale or McKelvie's continuous line. The final number for data analysis was slightly less than 75 (71) because some participants did not record or spoiled their second judgment.

Materials and Procedure

Participants made their initial judgments of 9 or 2143 following McKelvie's (2001) procedure, which was based on Birnbaum's (1999) internet experiment. With the numerical scale, they were given a slip of paper with the following question: "On a scale from 1 to 10, where 1 = very very small and 10 = very very large, please judge, how large is the number 9 (or 2143)?" When they had recorded their judgment, participants were asked to judge the second number (2143 after 9 or 9 after 2143). Finally, participants were asked if they had thought of the meaning of very very small or very very large when they were making their initial judgment. If they answered in the affirmative, they were asked to state what numbers they had in mind. With the continuous line, the procedure was the same except that the paper contained a horizontal 87 mm line with very very small written just beyond the end on the left and very very large written just beyond the end on the right. Participants placed a small vertical mark on the line to indicate their judgment. Subsequently, following the procedure outlined in Study 1, these ratings were measured in mm from the left hand side of the line and then converted to the numerical scale from 1 to 10, where 1 corresponded to the extreme left hand side of the line (0) and 10 corresponded to the extreme right hand side of the line (87 mm).

Results

Study 2A

For Study 2A, in which 9 and 2143 were compared using the numerical scale, a 2 X 2 (Order X Judgment) mixed ANOVA with repeated measure on judgment was conducted on the size ratings. The two orders were 9 then 2143 or 2143 then 9, and the two judgments were the ratings made first or second. Two effects were significant: order, $F(1, 34) = 12.46$, $p = .001$, $\eta 2 = .074$, and the interaction between order and judgment, $F(1, 34) = 49.87$, $p < .001$, $\eta 2 = .585$. The effect of judgment was not significant, $F(1, 34) = 1.37$, $p = .25$, $\eta 2 = .016$. Generally, ratings were higher when 9 was rated first than when 2143 was rated first. For the judgments made first, a post hoc t-test showed that the difference in rated size between 9 and 2143 was not significant (see Table 5), $t(34) = 0.32$, $p = .75$, d = 0.11. However, within-subjects, when 2143 was rated after 9, it was judged significantly larger than 9, $t(14) = 3.56$, $p = .003$, $d = 0.92$, and when 9 was rated after 2143, it was judged significantly smaller than 2143, $t(20) = 6.80$, $p < .001$, $d = 1.48$ (see Table 5).

JMIC scores were also calculated for each participant. In addition, 13 out of the 36 participants reported having had numbers in mind for very very small and very very large. As in Study 1, these reports for very very large were converted to a 5-point scale and correlated with JMIC scores to provide a validity check (see Table 3). The correlation for the 13 participants is positive (.955) and significant, and although sample sizes were small for each condition, both correlations were positive (.932, .730) and significant. Finally, and most importantly, an independent samples t-test showed that JMIC scores were lower for 9 than for 2143, $t(34) = 3.53$, $p = .001$, $d = 1.15$. These results are shown in Table 6.

Study 2B

For Study 2B (see Table 5), in which 9 and 2143 were compared using both the numerical scale and the continuous line, a 2 X 2 X 2 (Scale X Order X Judgment) mixed ANOVA showed two significant effects: order, $F(1, 67) = 18.03$, $p < .001$, $\eta 2 = .065$, and the interaction between the order and judgment, $F(1, 67) = 46.80$, $p < .001$, $\eta 2 = .396$. There were no significant effects of scale, $F(1, 67) = 0.90$, $p = .346$, $\eta 2 = .003$, nor of any of the interactions: scale X order, $F(1, 67) = 0.18$, $p = .68$, $\eta 2 = .0006$, scale X judgment, $F(1, 67) = 1.24$, $p = .269$, $\eta 2 = .011$, scale X order X judgment, $F(1, 67) = .075$, $p = 785$, $\eta 2 = .0006$. Generally, ratings were higher when 9 was rated first than when 2143 was rated first. Collapsing the data across the two scales and considering the judgments made first, a post hoc *t*-test showed that the difference in rated size between 9 and 2143 was not significant, $t(65) = 0.23$, $p = .82$, $d = .06$. When 2143 was rated after 9 within-subjects, it was judged significantly larger than 9, $t(37) = 7.67$, $p < .001$, $d = 1.24$, and when 9 was rated after 2143, it was judged significantly smaller than 2143, $t(32) = 2.96$, $p = .006$, $d = 0.52$.

The number of participants in each condition was planned on the basis of a large difference in perceived size between 9 and 2143. However, it is possible that the difference exists but is smaller than expected, which implies that sample size was too low to detect the effect. To investigate this possibility, the data from Study 2A,

which were all gathered with the numerical scale, were combined with the data from the two conditions in Study 2B in which that scale was employed. In the 2 X 2 (Order X Judgment) ANOVA, two effects were significant: the main effect of order, $F(1, 66) = 22.27$, $p < .001$, $\eta 2 = .072$, and the interaction between order and judgment, $F(1, 66) = 49.21$, $p < .001$, $\eta 2 = .475$. The effect of judgment was not significant, $F(1, 66) = 0.36$, $p = .552$, $\eta 2 = .0$. A post hoc *t*-test between the initial between-subjects ratings for 9 and 2143 was again not significant, $t(66) = 0.64$, $p = .53$, $d = 0.15$. However, within-subjects, when 2143 was rated after 9, it was rated as larger than 9, $t(28) = 5.92$, $p < .001$, $d = 1.10$, and when 9 was rated after 2143, it was rated as smaller than 2143, $t(38) = 4.47$, $p < .001$, $d = 0.72$.

For Study 2B, JMIC scores were calculated for each participant. In addition, 22 out of the 65 participants reported having had numbers in mind for very very small and very very large. As in Study 1, the reports for very very large were converted to a 5-point scale and correlated positively and significantly with JMIC scores. Because sample sizes were four or less in three of the four conditions, only a single correlation was calculated for all participants together. It was positive (.853) and significant (see Table 3). Finally, and most importantly, a 2 X 2 (Scale X Number) ANOVA showed that JMIC scores were lower for 9 than for 2143, $F(1, 67) = 18.0$, $p < .001$, $p = .01$, $\eta 2 = .209$ (see Table 6). Effect sizes were $d = 1.62$ for Birnbaum's numerical scale and $d = 1.02$ for the line but, in the absence of a significant interaction between scale and number, $F(1, 67) = 0.18$, $p = .676$, $\eta 2 = .002$, these effect sizes were not significantly different. Finally, the effect of scale, $F(1, 67) = 0.90$, $p = .345$ $\eta 2 = .010$, was not significant.

Meaning of Very Very Small and Very Very Large

As noted above, JMIC scores were validated against the reports for the meaning of very very large for people who answered this question after making their judgments. The present analysis considers these reports for participants in Studies 2A and 2B combined. As in the experiment by McKelvie (2001), the vast majority of participants in all conditions (30 out of 33 in total or 91%) reported that very very small was close to 1 (8 out of 8 [100%] after 9 was judged on the numerical scale, 7 out of 9 [78%] with 2143 on the numerical scale, 12 out of 13 [92%] with 9 on the line, and 3 out of 3 [100%] with 2143 on the line). In particular, when the number being judged was 2143, only 1 person (who was in the numerical scale condition) out of 12 (8%) reported that very very small was close to 1000, which is the lower bound of 4-digit numbers. For very very large, no participant (0 out of 12 [0%]) reported a value close to 10,000 when judging 2143; all numbers were much larger than that and were usually in the millions. When judging 9, the most popular response was also a very high number (at least 10,000 and in the millions), but 8 out of 21 (38%) people (5 out of 8 [63%] on Birnbaum's numerical scale and 4 out of 13 [31%] on the continuous line) reported that very very large was interpreted as 10, which is the first number above the single digits.

Reports were also classified along the same lines as McKelvie (2001), where the numbers were 9 and 221. When responses to very very small were classified as less than 9 or more than 9, the two people who chose more than 9 had judged 2143 on the numerical scale. In addition, when responses to very very large were classified as less than 9999 or more than 9999, the highest number of people for less than 9999 occurred with 9 on the numerical scale.

Discussion

Judgments of the Size of 9 and of 2143

In Study 2A with the numerical scale, and in Study 2B with both the numerical scale and the continuous line, participants who judged 9 and participants who judged 2143 rated them as similar in size. The nonsignificant effect on the continuous line is consistent with the nonsignificant difference between 9 and 221 on the same line reported by McKelvie (2001). McKelvie argued that this line, together with the verbal labels very very small and very very large, invoked a wide range of numbers, possibly all positive numbers. In this context, 9 and 221 were both small. Extending this line of argument, 2143 would also be relatively small in the context of a wide range of numbers.

However, it was stated above that the similar scores for 9 and for 221 (3.8, 3.5) on the line were higher than the value of 1 predicted by McKelvie's (2001) wide context spurious model, and that they could be better accounted for by the context size priming model. As a parallel to this, the similar values for 9 and 2143 (3.86, 3.88) on the line were higher than the value of 2 predicted by McKelvie. Consequently, they, too, can be better accounted for by the context size priming model than by McKelvie's wide context spurious model. That is, 9 invoked a smaller context than 2143, but both 9 and 2143 were judged to be similar in size relative to their own contexts.

Turning to the numerical scale, both McKelvie (2001) and Birnbaum (1999) found that 9 was rated as larger than 221. Birnbaum argued that this occurred because 9 invoked the context of single digit numbers and was rated relatively high and 221 invoked the context of triple digit numbers and was rated relatively low. In contrast, McKelvie criticized the idea that each number invoked its own context, suggesting that 221 was simply a relatively

low number in the context of a wide range of numbers and that the high rating of 9 occurred because participants interpreted very very large to literally mean 10. However, it was also suggested here that the numerical scale may have primed participants to invoke a different context for each number. Overall, it was argued above that the best account of these results is also provided by the context size priming model in which the difference between the contexts invoked by 9 and by 221 was greater on the numerical scale than on the line. Nevertheless, from any of these accounts, it would be expected that 9 would be also rated as greater than 2143 on the numerical scale.

The results were not in accord with this prediction. For people who judged 9 and for people who judged 2143, 9 was not rated as greater than 2143 on this scale in Studies 2A or 2B, even when the data were combined to increase sample sizes. If the present account of this result on the line is valid, it implies that 9 may have invoked a smaller context than 2143, but the difference was not greater on the numerical scale than on the line. That is, the numerical scale did not in this case serve the priming function of moving the invoked context of each number closer to single digits and quadruple digits. This explanation will be considered later when discussing the results for JMIC scores. If the present analysis is correct, the JMIC scores will be smaller for 9 than for 2143, and the effect will be similar on the numerical scale and on the line.

In contrast to the between-subjects comparisons, 9 was consistently judged to be smaller than 2143 when they were compared within-subjects. Effect sizes ranged from medium to very large. This is the same result obtained for 9 and for 221 by McKelvie (2001) and is consistent with Birnbaum's (1999) statement that number would be rated consistently with their size when the judgments were made by the same people.

Meaning of Very Very Small and Very Very Large

Of the participants who gave reports, the vast majority indicated that the meaning of very very small was close to 1, which is consistent with Birnbaum's (1999) specific context model for 9, but not for 2143, where 1000 should be the lower limit. Although the two people who chose a number more than 9 for very very small had judged 2143 on the numerical scale, only one included 1000 in their response. For very very large, when 9 had been judged, 10 was reported by 5 out of 8 (63%) people on the numerical scale and by 4 out of 13 (31%) people on the line. However, when 2143 had been judged, only one person gave a report (7000) somewhat close to the highest four-digit number (9999); most of the other 11 people reported that very very large exceeded one million.

Although only about one third of participants gave reports for the meaning of the labels, a low rate of response that was approximately half of the rate found by McKelvie (2001), these reports suggest that, for some participants on the numerical scale, and perhaps also on the line, 9 invoked the context of single digits, consistent with Birnbaum's model. For others who judged 9 on either scale, the context was a wide range of numbers, perhaps all numbers. However, for the vast majority of the participants who judged 2143 and who reported the meaning of very very small and very very large, the context was also a wide range of numbers. These reports are more consistent with McKelvie's (2001) suggestion that the labels very very small and very very large invoke a wide range of numbers. However, these results occurred with both scales, rather than simply on the line, as suggested by McKelvie (2001).

JMIC Scores

Firstly, the positive correlation between JMIC scores and the reported meaning of very very large replicates the correlation found from McKelvie's (2001) data and provides further support for the validity of the JMIC.

Secondly, in the analysis of JMIC scores taken from McKelvie's (2001) data, the inferred context for 9 was smaller than the inferred context for 221 on both the numerical scale and on the continuous line, but the size of the effect was less on the line compared to the numerical scale. In Study 2 here, the inferred context for 9 was again smaller than the inferred context for 2143. On the numerical scale, there were very large effect sizes of $d = 1.15$ in Study A and 1.62 in Study B, and on the continuous line scale, there was a very large effect size of 1.02. As with McKelvie's (2001) data, the effect size was again somewhat smaller on the line than on the numerical scale, but in this case the difference was not significant.

Although the direct reports for the meaning of very very small and very very large indicated that 9 might have invoked the context of single digit numbers for some participants, as held by Birnbaum's specific context model, the mean JMIC scores for the two numerical conditions (.37, .40) and for the line (.48) clearly exceed the value of zero predicted by this model and by the precise numerical priming model. In addition, although the reports also indicated that 2143 might have invoked a wide range of numbers, the mean JMIC scores for the numerical scale (.68, .69) and for the line (.72) are smaller than the value of .85 predicted by Birnbaum's specific context model, by McKelvie's wide context spurious model, and by the wide context specific priming model. The best account of the smaller JMIC scores for 9 than for 2143 on both scales is provided by the context size priming model. Indeed, when this model was applied to account for the similar judgments of 9 and 2143 on both scales, it was stated that it would be supported if the difference between JMIC scores for 9 and for 2143 was similar on both scales. The difference was indeed similar on both scales, supporting the context size priming model.

Of course, following the results of the analyses of McKelvie's (2001) experiment, it was predicted that the difference in the JMIC scores for 9 and for 2143 would be greater on the numerical scale than on the line because the numerical scale would prime the notion of contexts that approached the number of digits in each number, and that this would be accompanied by 9 being judged as greater than 2143. Clearly, the results for judgments and for JMIC scores did not accord with these predictions. One reason may be that this priming effect of the numerical scale decreases with larger numbers. More specifically, the effect may decline to the extent that for 2143 it is absent. This idea could be tested by obtaining judgments and JMIC scores for other numbers with steadily increasing digits.

General Discussion

The Juillet Measure of Inferred Context

One contribution of this paper was to propose the *Juillet Measure of Inferred Context* (JMIC). The purpose of this measure was to quantify and compare the contexts invoked by different numbers. The validity of the JMIC was supported by evidence that JMIC scores were positively correlated with direct reports of the meaning of the rating scale label "very very large" and it provided information that, in conjunction with the judgments themselves, helped to evaluate four models of number judgment (the specific context model, the wide context spurious model, the wide context specific priming model, and the context size priming model).

Comparison of Models

Which of the models best accounts for the number judgments and for the JMIC scores in this series of experiments (Birnbaum, 1999; McKelvie, 2001; Studies 1, 2A, and 2B here)?

Birnbaum (1999) proposed that, when judged for size, numbers induce and are evaluated within their own context, which is defined by how many digits there are in the number. The best evidence for this specific proposal occurred in some of the participant reports for the meaning of the scale labels very very small and very very large. In McKelvie's (2001) experiment, in which 9 and 221 were judged, most participants reported that they had been thinking of a number close to 1 as very very small. In addition, some participants reported that they had been thinking of 10 as a very very large number. In addition, for 221, a small number of participants reported that a very very small number was close to 100 and that a very very large number was close to 999. Similarly, in Study 2 here, when 9 was judged, most respondents stated that a number close to 1 was very very small and some participants, particularly on the numerical scale, reported that they had been thinking of 10 as a very very large number. On the other hand, for 2143, there was almost no evidence that participants thought of 1000 and 9999 as a very very small number or a very very large number, respectively.

These reports for 9 and for 221 are consistent with the precise requirements of Birnbaum's specific context model, but they constituted a minority of the reports, which were themselves given by only 63% of the participants. In addition, they were not clearly supported by the results for the number judgments or for the estimates of invoked context. Although some results (size judgments and JMIC scores for 9 vs. 221 on the numerical scale, JMIC scores on the line; JMIC scores for 9 vs. 2143) were in the direction predicted by Birnbaum's model, they deviated from the precise values that would be expected. Overall, the results are not in accord with Birnbaum's model.

McKelvie (2001) suggested that numbers do not induce their own context, but are simply judged in the context of a wide range of numbers. He also argued that the models are more fairly tested on the continuous line, because the numerical scale encouraged participants to spuriously interpret 10 as the very very high number, particularly when 9 was being judged. Many of the reports for the meaning of very very small and very very large are consistent with McKelvie's (2001) wide context spurious model because they ranged from 1 to numbers in the millions. Furthermore, on the line, 9 was judged to be similar in size to both 221 and 2134, which is reasonable if the ratings are low and the context is a wide range of numbers, perhaps all numbers. Furthermore, by this account, 9 was only judged as larger than 221 on the numerical scale because participants literally took the very very high number to be 10, not because the numbers induced their own particular digital context. Being close to 10, 9 was then rated as large. However, as with Birnbaum's (1999) specific context model, both the size judgments and the JMIC scores deviated from the precise values that would be expected from McKelvie's wide context spurious model. Moreover, because the wide context specific priming model implies that McKelvie's wide context notion applies on the line and Birnbaum's specific context notion model applies on the numerical scale, the size judgments and the JMIC scores also deviated from these specific predictions.

The best explanation of the results seems to be offered by the *context size priming model*, according to which contexts are induced by the target numbers, are positively associated with their size, and are primed by the numerical scale to be closer to Birnbaum's values. Each number is then judged for size within its own context. For 9 vs. 221, JMIC scores were smaller for 9 than for 221 on both rating scales, but the difference was greater on the numerical scale than on the line. Nine was judged to be larger than 221 on the numerical scale, but not on the line. It was argued that the numerical scale primed a context that was smaller on the numerical scale than on the line for 9

and larger on the numerical scale for 221. Consequently, on the numerical scale, 9 was larger in its context than 221 was in its context. For 9 vs. 2143, JMIC scores were smaller for 9 than for 2143 on both scales, but the difference was not greater on the numerical scale. As with 221, 9 was judged to be similar in size to 2143 on the line, but this was also true on the numerical scale. It was argued that the priming effect of the numerical scale on the context invoked by the numbers did not occur for 2143, leading to the same results on both scales for the size judgments. In general, it was suggested that the priming effect of the numerical scale is largest for a small number like 9 which is close to the numerical values on the scale (1, 10), but then decreases as numbers become larger, and disappears for 2143.

If this explanation is correct, the priming function of Birnbaum's (1999) numerical scale is interesting, but it adds a complication to the size judgment task. Moreover, the judgment of 9 on the numerical scale may be spurious if some participants interpret 10 as being the very very large number (McKelvie, 2001). Consequently, future research on this task should be conducted using the continuous line, or perhaps other kinds of scales that do not involve numbers. This would also contribute to the issue of rating scale format. For example, although the psychometric properties of the continuous line are similar to certain kinds of categorized rating scales, participants prefer the line, which also permits them to indicate how confident they feel in their judgment by indicating a range around it (McKelvie, 1978).

Finally, a more sensitive test of the various models might be obtained with different labels for the endpoints of the continuous line. "Very very small" and "very very large" imply an extremely wide range of numbers, possibly all numbers (McKelvie, 2001) and may be another complicating factor that affects judgments. In particular, the wide range of numbers invoked by these labels provides a context that may counteract the extent to which numbers induce their own contexts. One possibility would be to employ labels such as "low" and "high", which are less explicit in terms of their implied range. Such labels may allow more flexibility in their interpretation and permit contexts induced by the numbers themselves to exert their effects on size judgments. Indeed, some evidence already points in this direction. In contrast to McKelvie's (2001) finding that size judgments for 9 and 221 did not differ on the continuous line with the end labels "very very small" and "very very large", McKelvie (2012) found that 9 was judged to be larger than 221 on the continuous line with the end labels "low" and "high." However, as in the previous study, the size of the effect was smaller than on the numerical scale.

Another suggestion for future research is to increase sample size in the various experimental conditions. In Study 2, sample size was planned on the basis of a large effect size from previous work. However, if effects of number on invoked context are smaller, larger sample sizes will be required to detect them at the same level of power.

Between- vs. Within-Subjects Designs

The results also have implications for the methodological issue of between- vs. within-subjects designs, which has received considerable attention (Greenwald, 1976; Poulton, 1973). When people take part in more than one condition in a within-subjects design, contexts effects can be powerful and it is often suggested that results should be replicated between-subjects (Poulton, 1973). The present findings indicate that between-subjects design do not eliminate context effects. In addition, because the natural ordering of numbers was preserved within-subjects but not between-subjects, they add to previous evidence (e.g., Mahoney et al., 2011) that outcomes of experiments may differ according to which design is employed.

Conclusion and Implications

This investigation shows that number judgment can be biased by context and it suggests that contexts can be induced by the numbers themselves when they are rated by different people in a between-subjects design. It also introduces the JMIC as a measure of these contexts that, together with the results of the number judgments, provide support for the context size priming model of number judgment.

Although the present task does not seem immediately relevant to many situations in everyday life, numbers are sometimes judged. For example, price reductions were evaluated differently when they were framed in percentage terms than in dollar terms (Chen, Monroe, & Lou, 1998). In addition, changing a price from a round value to a value ending in 9c (e.g., $3.00 to $2.99) reduced perceived cost and increased sales (Manning & Sprott, 2009; Thomas & Morwitz, 2004). However, this occurred only if the left digit changed (e.g., not from $8.70 to $8.69). An interesting extension of the present work would be to ask whether there would be circumstances under which something priced at $9 might be seen as more expensive than if it were priced at $221.

Because the true answer was known in the present studies, the results reveal the precise degree to which invoked contexts can influence and distort judgments. In addition, the results might also apply in real-world conditions that resemble those in the present task, even if numbers are not judged directly (see Mook, 1983, for a similar argument). For example, as Birnbaum (1999) observed, if two different professors are judged for teaching competence by different groups of students, Professor A could be poorer than Professor B, yet may receive similar

or even higher mean ratings on competence than Professor B. Indeed, it was noted earlier in the paper that students did rate professors differently depending on the numerical context in which they made their judgments (Mellers & Birnbaum, 1983). A similar outcome can easily be envisaged in the medical system. For example, just as students may not judge professorial competence correctly, different people may render incorrect judgments about the treatment they received (Birnbaum, 1999) or about the physicians who delivered their treatment.

Finally, the results also provide information about the kinds of rating scales that can be used in studies of judgment. Given that the interpretation of the meaning of the numerical scale may be ambiguous in some circumstances, the present findings imply that the continuous line is preferable.

References

Birnbaum, M. H. (1974). Using contextual effects to derive psychophysical scales. *Perception & Psychophysics, 15*, 89-96.

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

Chen, S.-F. S., Monroe, K. B., & Lou, Y-C. (1998). The effects of framing price promotion messages on consumers' perceptions and purchase intentions. *Journal of Retailing, 74*, 353-372.

Choplin, J. M., & Logan, G. D. (2005). A memory-based account of automatic numerosity processing. *Memory & Cognition, 33*, 17-28.

Christensen, L. B., Johnson, R. B., & Turner, L. A. (2011). *Research methods, design, and analy*sis. Boston: Allyn & Bacon.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Rev. ed.). New York: Academic Press.

Dehaene, S., & Akhavein, R. (1995). Attention, automaticity, and levels or representation in number processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 314-326.

Epley, N., & Gilovich, T. G. (2006). The anchoring-and-adjustment heuristic. *Psychological Science, 17*, 311-318.

Galotti, K. M. (2004). *Cognitive psychology in and out of the Laboratory* (3rd ed.). Belmont, CA: Thomson Wadsworth.

Ganor-Stern, D. (2013). Are ½ and 0.5 represented in the same way? *Acta Psychologica, 142*, 299-307.

Greenwald, A. G. (1976). Within-subjects designs: To use or not to use? *Psychological Bulletin, 1976, 83*, 314-320.

Keren, G. B., & Raaijmakers, J. G. (1988). On between-subjects versus within-subjects comparisons in testing utility theory. *Organizational Behavior and Human Decision Processes, 41*, 233-247.

Lambdin, C., & Shaffer, V. A. (2009). Are within-subjects designs transparent? *Judgment and Decision Making, 4*, 554-566.

Lechelt, E. C. (1971). Spatial numerosity discrimination as contingent upon sensory and extrinsic factors. *Perception &Psychophysics, 10*, 180-184.

Mahoney, K. T., Buboltz, W., Levin, I. P., Doverspike, D., & Svyantek, D. J. (2011). Individual differences in a within-subjects risky choice framing study. *Personality and Individual Differences, 51*, 248-257.

Manning, K. C., & Sprott, D. E. (2009). Price endings, left-digit effects, and choice. *Journal of Consumer Research, 36*, 328-335.

McBurney, D. H. & White, T. L. (2004). *Research methods*, 6th ed. Belmont, CA: Wadsworth/Thomson.

McKelvie, S. J. (1978). Graphic rating scales – how many categories? *British Journal of Psychology, 69*, 185-202.

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(16),* 45-61.

McKelvie, S. J., & Shepley, K. (1977). Comparisons of intervals between subjective numbers, an extension. *Perceptual and Motor Skills, 45*, 1157-1158.

Mellers, B. A., & Birnbaum, M. H. (1983). Contextual effects in social judgment. *Journal of Experimental Social Psychology, 18*, 157-171.

Mellers, B. A., & Birnbaum, M. H. (1982). Loci of contextual effects in judgment. *Journal of Experimental Psychology: Human Perception and Performance, 8*, 582-601.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist, 38*, 379-387.

Poulton, E. C. (1973). Unwanted range effects from using within-subject experimental designs. *Psychological Bulletin, 80*, 113-121.

Randell, J. A. (2009, December). How accessibility might affect the Take the Best heuristic. *Journal of Scientific Psychology*, 30-33.

Rickard, T. C., Romero, S. G., Basso, G., Wharton, C., Flitman, S., & Grafman, J. (2000). The calculating brain: An fMRI study. *Neuropsychologia, 38*, 325-335.

Thomas, M., & Morwitz, V. (2004). Effects of framing on magnitude perceptions of prices. *Advances in Consumer Research, 31,* 454-456.
Thompson, C. A., & Opfer, J. E. (2010). How 15 hundred is like 15 cherries: Effect of progressive alignment on representational changes in numerical cognition. *Child Development, 81,* 1768-1786.
Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.
Vuokko, E., Niemivirta, M., & Hrlenius, P. (2013). Cortical activation patterns during subitizing and counting. *Brain Research, 1497,* 4-52.
Whalen, J., Gallistel, C. R., & Gelman, R. (1999). Nonverbal counting in humans: The psychophysics of number representation. *Psychological Science, 10,* 130-137.
Zhang, J. H., Walsh, C., & Bonnefon, J.-F. (2005). Between-subject or within-subject measures of regret: Dilemma and solution. *Journal of Experimental Social Psychology, 41,* 599-566.

Table 1

Predicted and Obtained Results for Number Judgments and for Scores on the Juillet Measure of Inferred Context (JMIC) (Study 1 with 9 and 221)

Scale	Model Predictions						Obtained Results		
	Birnbaum (1999)		McKelvie (2001)						
	Specific Context		Spurious		Specific Priming				
	9	221	9	221	9	221	9	221	
Number Judgments									
Numerical	10	2	9	1	10	2	5.1	3.1	Birnbaum
							6.6	3.2	McKelvie
Line	10	2	1	1	1	1	3.8	3.5	McKelvie
JMIC Scores									
Numerical	0	.85	.05	.90	0	.85	.36	.78	
Line	0	.85	.85	.90	.85	.90	.48	.72	

Notes. Minimum and maximum scores = 1, 10 (number judgments); 0, 0.90 (JMIC Scores).

Table 2

Examples of Possible Invoked Contexts for 9 and for 221, Judgments of the Sizes of the Numbers, and Resultant Values for the Juillet Measure of Inferred Context (JMIC)

Invoked Context	Order		JMIC	Order		JMIC
	First	Second		First	Second	
from first number	9	221		221	9	
1-10	8 or 9	9 or 10	.05 to .15	10	8 or 9	.05, .10
1-25	3 or 4	9 or 10	.30 to .40	10	3 or 4	.35, .40
1-50	2	9 or 10	.40, .45	10	2	.40
1-100	1	9 or 10	.45, .50	10	1	.45
1-500	1	4 or 5	.70, .75	4 or 5	1	.70, .75
1-1000	1	2 or 3	.80, .85	2 or 3	1	.80, .85
1-5000	1	1 or 2	.85, .90	1	1	.85, .90
1-10,000	1	1 or 2	.85, .90	1	1	.90
>10,000	1	1 or 2	.80, .90	1	1	.90

Notes. Order means presentation order (9 then 221 or 221 then 9); judgments are likely values on the scale from 1 to 10.

Table 3

Correlations between JMIC Scores and Reports of the Meaning of Very Very Large

Scale Condition	*n*	*r*	*p*
		Study 1	
All Participants	71	.658	< .001
Numerical: 9 then 221	11	.553	.078
Numerical: 221 then 9	21	.702	< .001
Line: 9 then 221	15	.676	.004
Line: 221 then 9	24	.473	.023
		Study 2A	
All Participants	13	.955	< .001
Numerical: 9 then 2143	5	.932	.021
Numerical: 221 then 2143	8	.730	.040
		Study 2B	
All Participants	22	.853	.001

Note. Reports for the meaning of very very large were scaled as follows: 1 = 1 to 10, 2 = 11 to 100, 3 = 101 to 1000, 4 = 1001 to 10,000, 5 = greater than 10,000.

Table 4

Predicted Results for Number Judgments and for Scores on the Juillet Measure of Inferred Context (JMIC) (Study 2 with 9 and 2143)

Scale	Model Predictions					
	Birnbaum (1999)		McKelvie (2001)			
	Specific Context		Spurious		Specific Priming	
	9	2143	9	2143	9	2143
Number Judgments						
Numerical	10	2	9	2	10	2
Line	10	2	1	2	1	2
JMIC Scores						
Numerical	0	.85	.05	.85	0	.85
Line	0	.85	.85	.85	.85	.85

Notes. Minimum and maximum scores = 1, 10 (number judgments), 0, 0.90 (JMIC Scores).

Table 5

Mean Ratings for the Size of 9 and of 2143 in Each Scale Condition in Studies 2A and 2B

Scale Condition	*n*	9		2143	
		M	*SD*	*M*	*SD*
Study 2A					
Numerical: 9 then 2143	15	5.13	3.52	7.40	3.45
Numerical: 2143 then 9	21	1.64	1.75	4.81	2.52
Study 2B					
Numerical: 9 then 2143	14	4.43	3.65	7.61	3.10
Numerical: 2143 then 9	18	2.50	3.17	3.74	2.71
Line: 9 then 2143	24	3.86	2.84	6.47	3.23
Line: 2143 then 9	15	1.70	1.59	3.88	2.37

Note. Minimum and maximum scores = 1, 10.

Table 6

JMIC Scores for 9 and 2143 in Each Scale Condition in Studies 2A and 2B

	9			2143		
	n	*M*	*SD*	*n*	*M*	*SD*
Study 2A						
Numerical	15	.37	.33	21	.68	.19
Study 2B						
Numerical	14	.40	.32	18	.69	.22
Line	24	.48	.28	15	.72	.17

Note. Minimum and maximum scores = 0, .90.