Exploring a Counterintuitive Finding with Methodological Implications

Why is 9 > 221 in a Between-subjects Design?

Stuart J. McKelvie
Department of Psychology
Bishop's University
2600 College Street, Sherbrooke, Québec J1M 1Z7, Canada.

Abstract

In the present experiment, 107 undergraduates judged the size of 9 or of 221, using either a numerical scale from 1 to 10 or a continuous line, both of which were anchored by the labels low and high. Replicating previous research with the numerical scale and more extreme labels (very very small, very very large), it was found that 9 was rated greater than 221. In addition, from reports of the meaning of the end points of the scales, it was concluded that 9 invoked a smaller context than 221, at least for some participants. These results add to concerns that between-subject designs are not free of context effects that may lead to anomalous outcomes.

Keywords: Between- subjects design, context effects, rating scales, number judgment

1. Introduction

In the social sciences in general, and in psychology in particular, groups are often compared on the judgments that they make. Two methodological issues that are important to this endeavour are the nature of the research design and the nature of the rating scale because both factors can have an impact on the results that are obtained.

1.1. Research Design

One issue that is frequently debated is whether to use a between-subjects design or a within-subjects designs (Greenwald, 1976). In the first case, where different people serve in the different experimental conditions, the key problem is the subject selection effect, in which the people in the various groups are not equated (Christensen, Johnson, & Turner, 2011, p. 184; McBurney & White, 2004, p. 175). The most common way of dealing with this problem is to allocate people randomly to the different conditions (Christensen et al., 2011, p. 202; , McBurney & white, 2004, p. 196). In the second case, where the same people serve in the different conditions, the key problem is the context effect, in which initial judgments may influence subsequent ones. The most common way to deal with this problem is to counterbalance the order of conditions (Christensen et al., 2011, p. 215; McBurney & white, 2004, pp. 270-271).

Unfortunately, counterbalancing does not always work, because carry-over effects may not be symmetrical in each direction (Poulton, 1973). Consequently, it has been suggested that sometimes a within-subjects design cannot be used and, if it can be used, must be supplemented by repeating the experiment with a between-subjects design (Poulton, 1973). However, in 1999, Birnbaum reported an anomolous result with a between-subjects design: when asked to judge how large a number was on a scale from 1 = very very small to 10 = very very large, people given 9 rated it higher than other people given 221. Birnbaum offered the explanation that 9 induced the context of single-digit numbers (1 to 9), within which it was rated relatively large, and 221 induced the context of triple-digit numbers (100 to 999), within which it was relatively small. That is, the between-subjects design is not free of context effects. He argued that care must be taken when interpreting results from experiments with between-subjects designs, and he pointed out some practical examples of how his experiment might serve as a model for judgments in everyday life. For example, if Dr. A. is rated high by one group of patients and Dr. B is rated low by another group of patients, it cannot be concluded that Dr. A. is better than Dr. B. In fact, Dr. A. could actually be worse. This kind of scenario is often encountered in the social sciences, where surveys are employed to obtain judgments from different groups of people.

However, McKelvie (2001) suggested that Birnbaum's findings might be a consequence of his numerical rating scale.

Specifically, McKelvie speculated that participants might have taken 10 to literally define what is meant by a very very large number (rather than taking 10 to represent or stand for a truly very very high number). If this were true, 9 would have been rated relatively high because it was close to 10, not because it was high in the context of single digits. McKelvie tested this hypothesis by replicating Birnbaum's experiment with Birnbaum's scale, but also by including a second rating scale in the form of a continuous line. It was bracketed by the same verbal labels very very small (on the left) and very very large (on the right), but did not have any numbers. If Birnbaum's result was replicated on this scale, it would eliminate McKelvie's objection that his result was a spurious consequence of the numerical scale, and would support Birnbaum's interpretation that 9 and 221 each invoked its own context.

On Birnbaum's numerical scale, McKelvie successfully replicated Birnbaum's finding that 9 was rated as greater than 221, and actually obtained a larger effect (standardized effect size d = 1.20 compared to Birnbaum's d = 0.71). However, his analysis of variance for the two scales with the two numbers showed a significant interaction between the variables, because the difference between the ratings of 9 and 221 e (d = 0.11) was not significant on the continuous line. McKelvie argued that his results showed that the higher rating of 9 over 221 on the numerical scale was spurious and that the nonsignificant effect on the line occurred because very very small and very very large induced a very wide context (perhaps all positive numbers), so that 9 and 221 were both rated as relatively small.

McKelvie supplemented his argument by reporting the results of a post-experimental question in which participants were asked if they had been thinking of the meaning of very very small and very very large when they were making their judgment and, if so, what the meaning was. For the 63% who answered in the affirmative, the vast majority in both conditions reported that very very small meant numbers close to 1. This is consistent with Birnbaum's hypothesis for 9, but not for 221, where 100 should be small. On the other hand, there was some support for Birnbaum's hypothesis for very very large: 17% (numerical scale) and 13% (line) of participants reported the meaning as 9 or 10 when 9 was judged, and 10% (numerical scale) and 20% (line) reported a number close to 999 when 221 was judged. However, most people stated that very very large meant numbers much greater than 1000, which is more consistent with McKelvie's idea that a very wide range of numbers was invoked on the line.

1.2. Rating Scales

In addition to the question of between- vs. within-subjects designs, there has also been considerable discussion about different kinds of rating scale, particularly their format (Madden & Bourdon, 1964). For example, should they be in the format of categories or a continuous line, and if the former, how many categories should there be (McKelvie, 1978)? Information gathered with rating scales depends on subjective judgment, and these judgments are relative: they must be made with reference to some context or standard (Parducci & Marshall, 1961). Most importantly for the present argument, there is often no objective standard according to which they are made (Fillenbaum, 1960). As part of this discussion, research has been conducted on the kind of label attached to the scales. In particular, it has been pointed out that the extremity of the labels creates a context in which ratings are made. For example, Wyatt and Meyers (1987) found that judgments were more widely spread when they were made in the context of labels that were less nearly absolute (e.g., very little, very much) compared to those that were more nearly absolute (e.g. completely false, completely true). According to Wyatt and Meyers, The more absolute labels create a context with more "psychological width". In another study, Lam and Stevens (1994) found that ratings were higher and more negatively skewed when they were obtained with the scale endpoints labelled strongly disagree to strongly agree compared to disagree to agree.

In the present case, as suggested by McKelvie (2001) when discussing his results with the continuous line, labelling the ends of the scale as very very small and very very large implies a wide range of numbers. It is possible that Birnbaum's hypothesis holds to some extent, as indicated by reports of a minority of McKelvie's participants (see above), but was overshadowed by the wide psychological context experienced by the majority. However, Birnbaum's hypothesis might be more likely to apply if there was more flexibility in the meaning of the labels. This might occur if the labels were less extreme.

1.3. The Present Experiment

The purpose of the present experiment was to examine this possibility by replicating McKelvie's experiment with the modification that the end points of the scale were labelled as low and high, which are less extreme than very very small and very very large.

It was predicted that 9 would be rated as greater than 221 on the numerical scale (perhaps because Birnbaum is correct or perhaps because the scale is spurious). However, if Birnbaum is correct, the statistical interaction between scale and number would not be significant because 9 should be rated as greater than 221 on both the numerical scale and on the continuous line.

2. Method

2.1. Participants

Sample size was planned as follows. Although McKelvie (2001) found an effect size of $d = 1.20$, it was decided that the effect to be detected should be Birnbaum's (1999) effect of $d = 0.71$. With alpha set a .05 and a power of .70, the required number in each condition was 25 (http://www.divms.uiowa.edu/~rlenth/Power/). Participants were 107 undergraduate university students who were assigned randomly to the four conditions with the proviso that there should be a similar number of females and males in each condition. Due to some inconsistencies in administration, there were slightly more people in the conditions in which 9 was rated. However, the numbers were similar for the numerical scale and the continuous line and the matching of females and males was preserved.

2.2. Materials and Procedure

The basic procedure was the similar to that adopted McKelvie (2001), which itself was modeled on Birnbaum's (1999) internet procedure, with the exception that participants were tested in person. For the numerical scale, participants read the following question: "On a scale from 1 to 10, where 1 = low and 10 = high, please judge, how large is the number 9 (or 221)"? After their judgment had been recorded, participants were asked if they had thought of the meaning of low or high when they were making their judgment. If they said that they had, they were asked what numbers they had in mind. With the continuous line was used, participants saw a horizontal 87 mm line with "low" written just beyond the end on the left and "high" written just beyond the end on the right. They made their rating of 9 or 221 by placing a small vertical mark on the line. These ratings were later measured in mm from the left hand side of the line and 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).

3. Results

3.1. Judgments of the Size of 9 and 221

Alpha was set at .05. The judgments were initially examined with a 2 X 2 X 2 (Scale X Number X Sex) factorial ANOVA. Because none of the effects of sex was significant, a 2 X 2 (Scale X Number) ANOVA was conducted. The main effects of scale, $F(1, 103) = 6.31$, $p = .014$, and of number, $F(1, 103) = 8.29$, $p = .005$, were significant. The interaction between the two variables was not significant ($p = .133$). From Table 1, it can be seen that ratings were generally higher on the numerical scale than on the continuous line (standardized effect size $d = 0.53$), and were also generally higher for 9 than for 221 ($d = 0.56$). Because there was particular interest in the comparison of 9 and 221 on each scale (replication of previous results on the numerical scale and investigation of the newly-labelled continuous line), the effect size was calculated for each one. The two results were 0.85 for the numerical scale and 0.28 for the line.

Table 1: Ratings of Size of Each Number on Each Scale

	9			221			Total		
Scale	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
Numerical	31	7.26	2.97	22	4.79	2.81	53	6.23	3.13
Line	30	4.99	3.32	24	4.22	2.14	54	4.65	2.86
Total	61	6.14	3.32	46	4.49	2.47			

Note. Minimum and maximum scores = 1, 10.

3.2. Meaning of Low and High

Of the 107 participants, 53 stated that they had thought of the meaning of low and 61 stated that they had thought of the meaning of high. In the first case, almost all reports (48 out of 53) were close to 1, and they were spread evenly among the four conditions. In the second case, reports were spread widely from single digits to infinity. Because almost all of the reports for low were similar and close to 1, the reports for high were taken as the measure of experienced context. For data analysis, these reports were classified as follows: 1 = less than or equal to 10, 2 = 11 to 100, 3 = 101 to 1000, 4 = 1001 to 10,000 and 5 = greater than 10,000. This coding system retains the importance of single digits and triple digits for Birnbaum's hypothesis.

These coded reports for the meaning of high were initially examined with a 2 X 2 X 2 (Scale X Number X Sex) ANOVA. Because sex was not significant, a 2 X 2 (Scale X Number) ANOVA was conducted, yielding one significant effect: number, $F(1, 57) = 29.59$, $p < .001$, The mean score for 9 was lower than the mean score for 221 ($d = 1.59$) (see Table 2).

Table 2: Reports of Meaning of High in Each Condition

	9			221		
Scale	*n*	*M*	*SD*	*n*	*M*	*SD*
Numerical	12	1.58	1.24	8	3.25	0.71
Line	24	1.75	1.22	17	3.24	0.66
Total	36	1.69	1.21	25	3.24	0.66

Note. Minimum and maximum scores = 1, 5.

Although the interaction between the two variables was not significant ($p = .755$), effect sizes were calculated for each scale because of the present interest in each of them. For the numerical scale, $d = 1.65$ and for the line, $d = 1.52$.

3.3. Meaning of Low and High and Judgments of the Size of 9 and 221

Secondly, correlation coefficients were calculated between these reports and the ratings of the size of the numbers. Taking all four conditions together, $r(N = 61) = -.684$, $p < .001$. the correlations in each of the four conditions were also negative, although not always significant due to small sample size: $r(N = 12) = -.919$, $p < .001$ for 9 on the numerical scale, $r(N = 8) = -.269$, $p = .520$ for 221 on the numerical scale, $r(N = 24) = -.747$, $p < .001$ for 9 on the line and $r(N = 17) = -.427$, $p = .087$ for 221 on the line.

To examine further the relationship between the coded report (representing the meaning of high) and the corresponding ratings of the size of 9 and 221, each report was classified as being consistent with Birnbaum's hypothesis or not. Reports were consistent if they were in the single digits (coded as 1) for the meaning of high when 9 was judged and if they were in the triple digits (coded as 3) for the meaning of high when 221 was judged. As a stimulus check, a 2 X 2 X 2 (Scale X Theory Consistent X Number) ANOVA was conducted on the reports for high. Because none of the effects of scale was significant, the analysis was repeated with a 2 X 2 (Theory Consistent X Number) ANOVA. Both main effect were significant: $F(1, 57) = 111.52$, $p < .001$ (theory consistent), and $F(1, 57) = 102.60$, $p < .001$ (number). By definition, the mean values of the theory-consistent reports for the meaning of high were 1.00 and 3.00 for 9 and 221 respectively. The non-theory-consistent reports were both higher (3.08 and 5.00 respectively).

The ratings of the size of 9 and 221 were then examined with a 2 X 2 X 2 (Scale X Theory Consistent X Number) ANOVA. Because none of the effects of scale were significant, a 2 X 2 (Theory Consistent X Number) ANOVA was conducted. Here, all three effects were significant: theory consistent, $F(1, 57) = 35.99$, $p < .001$, number, $F(1, 57) = 5.53$, $p = .022$, and the interaction, $F(1, 57) = 5.30$, $p = .025$. Table 3 shows that, generally, 9 was rated as greater than 221, $d = 0.57$, and ratings were higher when the coded reports were theory consistent than when they were not, $d = 1.95$. Post hoc *t*-tests showed 9 was rated as higher than 221 for theory consistent reports, $t(44) = 5.41$, $p < .001$, $d = 1.58$, but not for non-theory consistent reports, $d = 0.03$. Also, ratings for 9 were higher for theory consistent reports than for non-theory consistent reports, $t(34) = 9.80$, $p < .001$, $d = 3.38$. Ratings for 221

tended to be higher for theory consistent reports than for non-theory consistent reports, but this comparison only approached significance, $t(23) = 1.80$, $p = .086$, $d = 1.41$.

Table 3: Ratings of Size of Each Number for Theory Consistent and non Theory Consistent Reports of the Meaning of High

	9			221		
Consistent	*n*	*M*	*SD*	*n*	*M*	*SD*
Yes	24	8.23	1.70	22	4.76	2.59
No	12	2.05	1.95	3	2.01	0.95

Note. Minimum and maximum scores = 1, 10.

Finally, because not all participants stated that they had been thinking of the meaning of low and high when making their judgments, a 2 X 2 X 2 (Scale X Number X Judgment Report: Yes or No) ANOVA was conducted. The effect of judgment report was not significant and did not interact significantly with any of the other variables ($ps > .229$).

4. Discussion

The major result of the present experiment is that 9 was judged to be greater in size than 221. In addition, the lack of a significant interaction between scale and number indicates that this trend occurred on both Birnham's numerical scale and on the continuous line. The difference on the numerical scale replicates two previous findings (Birnbaum, 1999; McKelvie, 2001), and the effect size of 0.85, which is large by Cohen's (1977) guidelines of 0.20 for small, 0.50 for medium and 0.80 for large, falls in between the effects sizes in the previous experiments (0.71 and 1.20 respectively). Although the difference on the continuous line (0.28) was smaller than on the numerical scale, it is greater than the previous result with a line (0.11, McKelvie, 2001). In fact, the present results contrast with those of McKelvie (2001) because he found a significant interaction between scale and number, indicating a clear effect of number on the numerical scale and no effect on the line.

Overall, the present results are consistent with Birnbaum's (1999) hypothesis that 9 invokes the context of single digit numbers and 221 invokes the context of triple digit numbers. This hypothesis claims that 9 is rated as greater than 221 because it is relatively higher in its context than is 221 is in its context. The major reason why the present results differ from those of McKelvie (2001) is that they were obtained with the scales labelled as low to high rather than very very small to very very large. Low to high implies a smaller range of numbers than very very small to very very large, and is more ambiguous.

Under these conditions, participants may have been more likely to infer different contexts from the two numbers to be rated. Very very small and very very large implies a very wide context in which both numbers are relatively small. This interpretation is also consistent with other evidence that ambiguous scale labels may be given a different interpretation when different stimulus sets are judged in a between-subjects design (Schifferstein, Smeets, & Hallensleben, 2011).

At the same time, although the interaction between scale and number was not significant, the size of the difference on the line was smaller than on the numerical scale. This suggests that the difference on the line might be accounted for by the different contexts invoked by 9 and 221, whereas the larger difference on the numerical scale might be accounted for by the different contexts but also by the spurious effect of high being taken to literally mean 10, as suggested by McKelvie (2001) for very very large. It is also notable that the four mean scores in the present experiment (7.26 for 9 on the numerical scale, 4.79 for 221 on the numerical scale, 4.99 for 9 on the line and 4.22 for 221 on the line) are higher than the corresponding four mean scores found by McKelvie (6.58, 3.21, 3.80 and 3.52). This result is consistent with the notion that the range of numbers suggested by very very small to very very large is greater than the range suggested by low and high. As observed above, both 9 and 221 are relatively lower in the wider range. The result is also similar to the finding that ratings were higher when made in the context of less extreme labels (disagree, agree) than more extreme labels (strongly disagree, strongly agree) (Lam & Stevens, 1994).

The post-experimental reports of the meaning of high are also consistent with Birnbaum's hypothesis. The mean score for the code for the meaning of 9 (1.71) was lower than the mean score for the meaning of 221 (3.24), and the effect size of 1.59 for this difference was extremely large. Moreover, this difference was also extremely large on both scales (1.65, 1.52 for the numerical scale and the line respectively). In addition, the highly significant correlation of -.684 between these reports and the number ratings shows that a narrower context was associated with a higher rating and a wider context was associated with a lower rating. This implies that the ratings of the size of 9 were higher than the ratings of the size of 221 because 9 was relatively higher in its smaller context and 221 was relatively lower in its larger context, as hypothesized by Birnbaum (1999).

However, according to a strict interpretation of Birnbaum's hypothesis, 9 is the highest of the single digits, which implies that its mean rating should be 10. Similarly, 221 lies .1346 of the distance between 100 and 999. On the scale from 1 to 10, this would mean that 221 would receive a rating of 1 + (9 X .1346) = 2.21, which could be 2 or 3. The mean rating for 9 (6.14) is below 10 and the mean rating for 221 (4.49) is above 3. This implies that Birnbaum's hypothesis does not hold in its strictest form. Rather, it may be the case that 9 invokes a smaller context than 221, but not specifically single digits or triple digits respectively.

Another possibility is that Birnbaum's hypothesis applies to some people but not to others. Some evidence to support this idea can be found in the analysis in which the ratings of the size of 9 and 221 were related to the question of whether or not the reports of the meaning of high were consistent with Birnbaum's hypothesis. In some cases they were and in some cases they were not. Moreover, when they were consistent with Birnbaum's hypothesis, the size of 9 was rated as much larger than the size of 221 (d = 1.58), and when they were not, this difference was almost zero (d = 0.03). In addition, the two theory-consistent ratings (8.23 for 9 and 4.76 for 221) are significantly higher than the two non-theory-consistent ratings (2.05 for 9 and 2.01 for 221). Because the two smaller ratings were given by people who reported much higher numbers for the meaning of high, they were made in a wider context. This is the context that McKelvie (2001) suggested was invoked by the labels very very small to very very large. In other words, it seems that Birnbaum's hypothesis applies to some, even most (see Table 3), participants, but not to others. For these people, "low" and "high" invoked a wide context.

One weakness in the analysis of the reports for the meaning of low and high is that not all participants provided them. This could mean that the others did not think of the meaning of the labels, did not report their thoughts or, less likely given the time gap between making the ratings and giving their reports, had forgotten their thoughts. However, when the independent variable of whether or not a report was given was included in the analysis of the relationship between scale and number on the one hand and ratings on the other, it did not interact significantly with any of the variables. That is, the same pattern of results for the ratings of the size of 9 and 221 was obtained for people who did and who did not give reports of the meaning of high.

5. Conclusion

Two important methodological issues in the social sciences and in psychology were examined in this paper: the kind of research design and the kind of rating scale. The starting point was Birnbaum's (1999) anomalous finding that 9 was rated as greater than 221 in a between-subjects design. He argued that 9 invoked that context of single-digit numbers and was rated as relatively large, whereas 221 invoked the context of triple-digit numbers and was rated as relatively small. The present experiment provides evidence that supports a slightly-modified version of the first part of Birnbaum's (1999) hypothesis: 9 invokes a smaller context than 221 for some people but not for others. However, there was direct evidence of a negative relationship between the range of the invoked context and the judgments of number size, which supports the second part of Birnbaum's hypothesis.

Consequently, the present research adds to the literature on the use of between-subjects designs and shows that, under certain circumstances, the results can be misleading. In addition, because Birnbaum's numerical scale may lead to spurious ratings, future research on the effects of context in a between-subjects design should be conducted with the continuous line. One interesting extension of research would be to discover conditions in everyday life that that are simulated by the present task, and investigate if similar results would be obtained.

References

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

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.

Fillenbaum, S. (1960). The effect of distributional skewing upon judgment with free choice of scale. *The American Journal of Psychology, 73*, 132-136.

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

Lam, T. C., & Stevens, J. J. (1994). Effects of content polarization, item wording, and rating scale width on rating response. *Applied Measurement in Education, 7*, 141-158.

Madden, J. M., & Bourdon, R. D. (1964). Effects of variations in rating scale format on judgment. *Journal of Applied Psychology, 48*, 147-151.

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

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

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

Parducci, A., & Marshall, L. M. (1961). Context-effects in judgments of length. *The American Journal of Psychology, 74*, 576-583.

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

Shifferstein, H. N. J., Smeets, M. A. M., & Hallensleben, R. (2011). Stimulus sets can induce in descriptor meanings in product evaluation tasks. *Acta Psychologica, 138*, 237-243.

Wyatt, R. C., & Meyers, L. S. (1987). Psychometric properties of four 5-point Likert type response scales. *Educational and Psychological Measurement, 1987, 47*, 27-35.