Judgment and Decision Making, Vol. 4, No. 7, December 2009, pp. 554–566

Are within-subjects designs transparent?

Charles Lambdin* and Victoria A. Shaffer
Department of Psychology, Wichita State University

Abstract

Researchers frequently argue that within-subjects designs should be avoided because they result in research hypotheses that are transparent to the subjects in the study. This conjecture was empirically tested by replicating several classic between-subjects experiments as within-subjects designs. In two additional experiments, psychology students were given the within-subjects versions of these studies and asked to guess what the researcher was hoping to find (i.e. the research hypothesis), and members of the Society for Judgment and Decision Making (SJDM) were asked to predict how well students would perform this task. On the whole, students were unable to identify the research hypothesis when provided with the within-subjects version of the experiments. Furthermore, SJDM members were largely inaccurate in their predictions of the transparency of a within-subjects design.

Keywords: methodology, research design.

1 Introduction

In the field of psychology, there is a long-standing controversy over the appropriate use of between- and within-subjects designs. Within-subjects designs have greater power and less variability, but many researchers eschew their use for two reasons. First, it has been argued that within-subjects designs render our research hypotheses transparent (e.g., Tversky & Kahneman, 1983). That is, in a within-subjects design, subjects will be aware of the purposes of our experiment and may behave accordingly, thus posing a threat to the internal validity of the experiment. Second, some have argued that life is more similar to a between-subjects design (Fischhoff, Slovic & Lichtenstein, 1979; Kahneman, Slovic & Tversky, 1982). Therefore, between-subjects designs increase the generalizability of the experimental findings. However, others have argued that between-subjects designs pose their own risks. Parducci (1965) and Birnbaum (1999) contend that, particularly with subjective judgments, the between-subjects design should be abandoned because it results in the confounding of context and stimulus.

The most famous demonstration of this principle was provided by Birnbaum (1999), who showed that in a between-subjects design subjects rated the number 9 as being significantly larger than 221. Theoretically, in a between-subjects design, the two conditions are identical except for the manipulation of a single stimulus. In this case, the stimulus to be manipulated was the number being rated, 9 or 221. However, Birnbaum argues that subjective judgments cannot be made in isolation; they require a context. In a within-subjects design, the context is specified; the two (or more) conditions are compared to each other. In a between-subjects design, the subjects are left to construct their own context to evaluate the stimulus. When this is the case, it is very likely that different contexts will be invoked for different stimuli. In this example, 9 is likely to bring to mind other single-digit numbers for comparison, thus leaving the impression that 9 is a relatively large number. In contrast, 221 is likely to bring to mind other triple-digit numbers for comparison, leaving the impression that 221 is a relatively small number. Thus, in a between-subjects design both the stimuli and the context vary between conditions, confounding the results.

Although the relative merits may be theoretically debated, what is more troubling is that hypotheses tested in these two designs often do not result in the same conclusions (Grice, 1966). For example, in between-subjects comparisons, manipulations of base rates do not affect judgments of probability, leading to the conclusion that base rates are ignored (e.g., Kahneman & Tversky, 1973). However, in within-subjects comparisons, base rates have large, significant effects on judgments of probability, leading to the conclusion that base rates are *not* ignored (Birnbaum & Mellers, 1983).

Despite the concerns about using between-subjects designs to evaluate subjective judgments raised by Birnbaum and others, it appears that the within-subjects design has fallen out of favor in many areas of psychology. In particular, within judgment and decision mak-

*These experiments are based on a dissertation submitted by the first author in partial fulfillment of the requirements for a Ph.D. degree at Wichita State University. Address: Victoria A. Shaffer, Department of Psychology, Wichita State University, 1845 Fairmount St., Wichita, KS, 67260–0034. E-mail: victoria.shaffer@wichita.edu

ing there exist many findings supported almost exclusively by between-subjects data. Examples include the hindsight bias (Fischhoff, 1975), research on reason-based choice (Shafir, 1993), availability (Schwarz, Bless, Strack, Klumpp, Rittenauer-Schatka & Simons, 1991), support theory (Tversky & Koehler, 1994) and many classic demonstrations of heuristics and biases (e.g., Kahneman et al., 1982). Researchers in these areas have frequently argued that between-subjects designs are more appropriate. One of the main reasons cited for the superiority of between-subjects designs is the belief that within-subjects designs are transparent (Bastardi & Shafir, 1998; Fischhoff, Slovic & Lichtenstein, 1979; Kahneman & Frederick, 2005; Tversky & Kahneman, 1983). Although this appears to be a popular reason for rejecting the within-subjects design, this assertion has never been empirically tested.

Thus, this gap in the literature inspired the two specific aims of this research. The first aim was to determine if classic examples of between-subjects designs could be replicated using within-subjects designs. The second was to empirically test the hypothesis that within-subjects designs are transparent.

2 Experiment 1

In Experiment 1, we attempted to replicate three classic between-subjects designs (Shafir, 1993; Tversky & Kahneman, 1981; Tversky & Kahneman, 1986) in a within-subjects format. In Shafir (1993), subjects were asked to pretend they were serving on a jury for a custody trial. They were presented with two generic parents, Parent *A* and Parent *B*. Parent *A* had a number of average attributes, whereas Parent *B* had both very positive and negative attributes. Shafir hypothesized that, because of the theory of reason-based choice, when asked to which parent custody should be awarded, subjects would selectively search for reasons to award custody. Since Parent *B* has more extreme positive attributes than *A*, most subjects would award custody to Parent *B*. Similarly, when asked to which parent custody should be denied, subjects would selectively search for reasons to deny custody. Since Parent *B* has more extreme negative attributes than *A*, most subjects would also deny custody to *B*.

In the Asian disease problem, Tversky and Kahneman (1986) presented subjects with two "programs" which were proposed solutions to a hypothetical outbreak of an "Asian disease", of which 600 were expected to die. One group of subjects was given a choice between Programs *A* and *B* and another group was given a choice between Programs *C* and *D*. Programs *A* and *B* were logically equivalent to Programs *C* and *D*. However, Programs *A* and *B* were presented in terms of lives saved (the "survival" frame) and Programs *C* and *D* were presented in terms of lives lost (the "mortality" frame). For instance, Program *A* states that, if adopted, 200 people will be saved. Program *C* states that, if adopted, 400 people will certainly die. Since the sample space is 600 people, these two statements should be seen as imparting the same information. Tversky and Kahneman (1981) hypothesized that a preference reversal would occur between the two frames, a framing effect. Subjects would choose Program *A* in the first pair and Program *D* in the second pair. Framing effects occur when different descriptions of functionally equivalent information cause people's preferences to differ.

In Tversky and Kahneman's (1986) marbles lotteries experiment, subjects were asked to choose one of two lotteries that they would like to play; these "lotteries" involved drawing a colored marble from a jar. One group of subjects was given Options *A* and *B*, and another group of subjects was shown Options *C* and *D*, both given below.

Option A				
90% white	6% red	1% green	1% blue	2% yellow
\$0	win \$45	win \$30	lose \$15	lose \$15
Option B				
90% white	6% red	1% green	1% blue	2% yellow
\$0	win \$45	win \$45	lose \$10	lose \$15

Option C			
90% white	6% red	1% green	3% yellow
\$0	win \$45	win \$30	lose \$15
Option D			
90% white	7% red	1% green	2% yellow
\$0	win \$45	lose \$10	lose \$15

Between Options *A* and *B*, *B* is the dominating lottery. The dominating lottery is the one with the better odds of winning. Between Options *C* and *D*, *D* is the dominating lottery. Options *A* and *B* are the same lotteries as Options *C* and *D*. The 6% red and 1% green to win \$45 in Option *B* have simply been combined into Option *D's* 7% red to win \$45. Similarly, Option *A*'s 1% blue and 2% yellow to lose \$15 have been combined into Option *C*'s 3% yellow to lose \$15. Because of this, Option *D* now has two losing outcomes and only one winning, whereas Option *C* now has two winning outcomes and only one losing. It was hypothesized that subjects would choose Option *C* over *D*, even though Option *D* is the same as *B* and dominates *C*.

2.1 Method

Eighty-nine undergraduate students at Wichita State University volunteered to participate in a brief survey during a psychology class. All students received extra credit for their participation. The survey contained a within-subjects adaptation of three between-subjects experiments: the child custody case (Shafir, 1993), the Asian disease problem (Tversky & Kahneman, 1981) and the marbles lotteries (Tversky & Kahneman, 1986). See Appendix A for the full text of all three experiments. The order in which the two conditions were presented was randomized for each design. Because the order of presentation is randomized, the use of a within-subjects design allows for both between- and within-subjects comparisons of the same data. In each of the three studies there were two conditions (*A* and *B*). Half of the subjects responded to *A* then *B,* and the other half responded to *B* then *A*. If only the first response is analyzed, then between-subjects comparisons can be employed, albeit with only half the sample size.

Additionally, although we are discussing the two designs as completely different methods, they are really best thought of as two endpoints on a single continuum. A within-subjects design, in its most pure form, would provide all of the conditions to the subjects simultaneously. Alternatively, you could have the different conditions on different pages, or administer the different conditions on different days, weeks or months. As illustrated, these designs successively move further away from a pure within-subjects design toward the between-subjects endpoint. To test whether the place on this continuum matters, half of the subjects saw the two conditions presented on the same page and half saw the two conditions on two separate pages.

2.2 Results and discussion

2.2.1 Testing for differences along the "within-subjects continuum"

The purpose of the replications was to see whether the overall patterns of data would be similar between designs. The point, therefore, of testing for differences along the within-subjects continuum is not to altogether rule out their presence, but rather to ensure that — if present — they are not of a magnitude that would create a qualitative change in the overall pattern of data. With this study, we had adequate power (i.e., 80%) to detect differences of 20 percentage points between conditions presented on the same page and conditions presented on separate pages. We found no significant differences between the results of these three within-subjects studies presented on the same vs. different pages. Therefore, the results reported below are collapsed across this condition.

Table 1: Percent (*N*) of subjects choosing Parent *A*.

	Award	Deny
Within-subjects	74% (66)	24% (21)
Between-subjects	73% (33)	23% (10)

Table 2: Percent (*N*) of subjects choosing the risk-averse option

	Survival frame	Mortality frame
Within-subjects	62% (55)	34% (30)
Between-subjects	76% (34)	27% (12)

The child custody case: The data presented in Table 1 show the percentage of subjects choosing Parent *A* in the award and deny conditions.

The within- and between-subjects analyses yielded identical conclusions. In both the "award" and "deny" conditions, the majority of subjects indicated that Parent *A* should have custody. These data did not replicate Shafir's (1993) original findings that people will award and deny custody to the same parent due to reason-based choice.

The Asian disease problem The data presented in Table 2 show the percentage of subjects choosing the risk-averse option in the survival and mortality frames.

In both the within- and between-subjects analyses of this data, the majority of subjects chose the risk-averse option in the survival frame and the risk-seeking option in the mortality frame. This replicates Tversky and Kahneman's (1981) original results.

Marbles lotteries The data presented in Table 3 show the percentage of subjects choosing the dominating option in pair one (Option *A* vs. Option *B*) and pair two (Option *C* vs. Option *D*). In pair 1, Option *B* is the dominant option, hence the rational choice. In pair 2, Option *D* is the dominant option. The within-subjects analyses used data from both pairs for all 89 subjects. The between-subjects analyses used data from only the first pair shown. Thus, half of the subjects ($N = 45$) saw Pair 1 first and half of the subjects ($N = 44$) saw Pair 2 first.

For both the within- and between-subjects analyses, the majority of subjects chose the dominating lottery in Pair 1 but not in Pair 2. This replicates Kahneman and Tversky's (1986) original findings.

In Experiment 1, three famous between-subjects experiments were replicated using a within-subjects design to determine the impact of design type on the findings. We

Table 3: Percent (*N*) of subjects choosing the dominant lottery.

	Pair 1	Pair 2
Within-subjects	96% (85)	33% (29)
Between-subjects	96% (43)	39% (17)

were able to replicate the original findings in two of the three experiments. In addition, the between- and within-subjects analyses of the same data led to the same conclusions in all cases. Experiment 1 supports the conclusion that between- and within-subjects designs may be more interchangeable than many researchers think. And, there may not be a need to dismiss the within-subjects design *a priori* on the grounds that it will decrease internal validity.

3 Experiment 2

In this experiment, we directly tested the assertion that within-subjects designs are more transparent than between-subjects designs (e.g., Bastardi & Shafir, 1998; Fischhoff, Slovic & Lichtenstein, 1979; Kahneman & Frederick, 2005). To do so, we presented psychology students with the within-subjects versions of three between-subjects designs (the same three studies used in Experiment 1): the child custody case (Shafir, 1993), the Asian disease problem (Tverksy & Kahneman, 1981) and the marbles lotteries (Tversky & Kahneman, 1986). Students were asked to figure out what the experimenter was looking for (i.e., identify the research hypothesis). These three studies were chosen because we predicted that they would vary in their transparency to undergraduate students. Specifically, we predicted that the child custody case would be most transparent and the marbles lotteries would be the least transparent. Thus, we believed that most psychology students would be able to guess Shafir's research hypothesis in the child custody case. However, we believed that Tverksy and Kahneman's hypothesis in the marbles lotteries would be opaque to the students.

3.1 Method

Eighty undergraduate students at Wichita State University volunteered to participate for extra credit. Eight subjects were dropped because they did not answer all of the questions, resulting in a sample size of 72. Before beginning the experiment, students were given a brief tutorial describing the concept of a hypothesis and illustrating how researchers design experiments to test their hypotheses. In addition, they were asked to identify the research hypothesis of a fictitious experiment, which was designed to be extremely transparent. This fictitious experiment served as a manipulation check to make sure that all subjects understood their task and could identify a hypothesis in a very simple research design. See Appendix B for the full text of the instructions and manipulation check. Subjects were then presented with the within-subjects adaptations of the child custody case, the Asian disease problem and the marbles lotteries (the same materials presented in Experiment 1). Half of the subjects saw the two within-subjects conditions on the same page and half saw the two conditions on different pages. After completing each study, subjects were asked to describe the experimenter's hypothesis, rate their confidence in their response (on a 7-point Likert scale) and rate the transparency of the design as completely transparent, somewhat transparent or not transparent at all.

Subjects' descriptions of the research hypotheses were judged to be correct or incorrect. The criteria upon which the responses were judged were very lenient. For example, in the child custody case, subjects merely needed to mention that changing the phrasing or wording would in some way impact the results. Thus, exhibiting some minor amount of insight was considered a successful completion of the task.

3.2 Results and discussion

As in Experiment 1, there were no significant differences between the results of the three studies when the conditions were presented on the same page vs. different pages. Thus, the results reported have been collapsed across this condition.

The manipulation check indicated that the vast majority of subjects understood the instructions and were able to recognize and articulate the hypothesis of a simple research design. However, nine of the 72 subjects did not pass the manipulation check. Analyses were done both with and without these nine subjects. The conclusions remained the same; therefore, the results below include all 72 subjects.

Although most of the subjects were able to identify the research hypothesis in the manipulation check, they were generally unable to do so with any of the three published experiments. Only 7% of subjects were able to correctly articulate some portion of the research hypothesis in the child custody case, the study deemed to be the most transparent *a priori* (95% *CI*: .01 to .13). Thirty-two percent of subjects correctly identified the research hypothesis in the Asian disease problem (95% *CI*: .21 to .43), while only 3% of subjects correctly identified the hypothesis in the marbles lotteries (95% *CI*: .00 to .07).

Subjects, therefore, were most accurate with the Asian disease problem, though it should perhaps be reiterated

Table 4: Confidence in ability to guess research hypothesis.

	M (*SD*)	95% *CI*
Manipulation check	2.32 (1.12)	2.06 to 2.58
Child custody case	3.32 (1.46)	2.98 to 3.66
Asian disease problem	3.01 (1.48)	2.67 to 3.35
Marbles lotteries	3.72 (1.65)	3.34 to 4.10

that these figures were arrived at by being very lenient with the criteria for success. Examples of subjects' guesses that were counted as correctly identifying the research hypothesis for the Asian disease problem include: "If rewording the question makes a difference in the choice" and "Changing the wording of the questionnaire from life to death expectancies will influence responses." Examples of guesses judged incorrect include: "The value of other people's lives," "Maybe how we let figures impress us," "What is considered a 'loss' of population," "It is better to take a risk than leave hundreds helpless" and "How many people will die based on the program chosen." This last quote is of a subject who was very confident he was correct and who rated the scenario as being "completely transparent."

Although this appeared to be a particularly difficult task for most subjects, they reported being somewhat confident in their responses. Table 4 lists the means and standard deviations for the confidence ratings for the manipulation check and the three studies (1 = "extremely confident" and 7 = "extremely unconfident").

Subjects were most confident in their responses to the manipulation check, followed by the Asian disease problem, the child custody case and the marbles lotteries. The ordinal relationship between the confidence estimates matches subjects' accuracy; subjects were most accurate (and were most confident) with the manipulation check and least accurate (and least confident) with the marbles lotteries. In contrast, subjects' categorization of transparency did not appear to be very sensitive with respect to their accuracy.

Given that 88% of subjects accurately described the research hypothesis in the manipulation check, this study should have been characterized as completely transparent by more than 36% of the subjects. Furthermore, a quarter of the subjects categorized the child custody case and the Asian disease problem as completely transparent. Given that very few people accurately identified the research hypotheses in these cases, this categorization seems largely optimistic. See Table 5 for transparency ratings across all four studies.

Experiment 2 directly tested the claim that within-subjects designs make research hypotheses transparent to subjects, thus potentially increasing demand characteristics and reducing internal validity. Taken together, both Experiments 1 and 2 provide evidence that within-subjects designs are largely opaque. The vast majority of psychology students were unable to identify the research hypothesis from within-subjects adaptations of three famous between-subjects experiments. Additionally, the psychology students were essentially unaware of their inability to do so. Subjects tended to be a least somewhat confident in their responses and only a small minority of subjects categorized these experiments as "not transparent".

Table 5: Ratings of transparency by undergraduate students, (*N*).

	Completely transparent	Somewhat transparent	Not transparent
Manipulation check	36% (26)	60% (43)	4% (3)
Marbles lotteries	17% (12)	53% (38)	31% (22)
Child custody case	25% (18)	51% (37)	24% (17)
Asian disease problem	25% (18)	60% (43)	15% (11)

4 Experiment 3

This final experiment was designed to determine if researchers are accurately able to predict the transparency of within-subjects research designs. To do so, we asked members of the Society for Judgment and Decision Making, who should be very familiar with all three studies, to categorize these studies as "completely transparent", "somewhat transparent" or "not transparent at all".

4.1 Method

Participation was solicited from the members of the Society for Judgment and Decision Making via the society's mailing list. Forty-eight members began the online survey; two subjects were removed from the data because they answered only a couple of questions.

4.2 Results and discussion

Members of the Society for Judgment and Decision Making (SJDM) shared our intuition that the child custody case would be the most transparent and the marbles lot-

Table 6: Ratings of transparency by members of SJDM, % (*N*).

	Completely transparent	Somewhat transparent	Not transparent
Child custody case	54% (25)	33% (15)	13% (6)
Asian disease problem	17% (8)	67% (31)	15% (7)
Marbles lotteries	4% (2)	37% (17)	59% (27)

teries would be the least transparent; see Table 6 for the transparency ratings of the three studies.

Although there was considerable agreement in the predicted transparency of the studies, these predictions appear to be largely inaccurate. Recall from Experiment 2 that undergraduate students performed poorly on this task across all three studies. Subjects were essentially unable to identify the research hypothesis in any of the three studies, the only exception being a sizeable minority of students (32%) who were able to identify what was being manipulated in the Asian disease problem. Thus, although SJDM members predicted variability in the transparency of these studies, undergraduate students had very little success identifying the research hypothesis in any of the studies.

In addition, the transparency ratings provided by the SJDM members can be compared with the transparency ratings provided by the undergraduate students. The two groups differed in their assessment of the transparency of the marbles lotteries, χ^2 (2, N = 118) = 10.45, $p <$ .05. The majority of SJDM members (59%) thought that the marbles lotteries would not be transparent. The majority of undergraduates (70%), however, rated the marbles lotteries as being either "somewhat" or "completely transparent". The two groups also differed in their transparecy ratings of the child custody case, χ^2 (2, N = 118) = 10.11, $p < .05$. A majority of SJDM members (54%) thought the child custody case experiment would be "completely transparent", whereas a majority of undergraduates (51%) rated it as being "somewhat transparent". However, the two groups did not differ in their transparency ratings of the Asian disease problem, χ^2 (2, N = 118) = 1.00, $p > .05$. The majority of both groups rated this design as "somewhat transparent".

Although ratings of transparency differed between undergraduate students and SJDM members (who are very familiar with these three studies), the SJDM members were no more accurate than the undergraduate students at assessing the transparency of within-subjects research designs. SJDM members erroneously believed that subjects would find some within-subjects research designs to be transparent when, in reality, very few undergraduate students were accurately able to describe the research hypothesis when presented with all conditions of a research design.

5 General Discussion

Within-subjects designs have a number of well-documented benefits; most importantly, they increase the power of an experiment (Kirk, 1995; Zimmerman, 1997) and avoid stimulus and context confounds (Birbaum, 1999). Researchers, however, often abandon within-subjects designs amid *a priori* concerns that increased task transparency will alter experimental outcomes. In this paper, we tried to assuage these concerns by replicating three famous between-subjects studies in a within-subjects format and empirically testing the transparency of the within-subjects designs. In Experiment 1, we were successfully able to replicate the findings in two of the three studies (the Asian disease problem and the marbles lotteries). Although our conclusions did not match those of Shafir (1993), both the between- and within-subjects analyses yielded identical results. Thus, we argue that this finding may not easily replicate with either design.

In Experiments 2 and 3, undergraduate students and members of the Society for Judgment and Decision Making categorized these within-subjects adaptations on their transparency: "completely transparent", "somewhat transparent" or "not transparent at all". Although these two groups differed in their assessment of the transparency of within-subjects designs, neither assessment proved to be accurate. Both experienced researchers and undergraduate students overestimated the transparency of these within-subjects designs. Very few undergraduate students demonstrated any ability to identify the hypotheses driving these original experiments. Even the research hypothesis that most agreed would be readily apparent was not obvious to a vast majority of undergraduate students. A limitation of these studies is the small sample sizes, both of subjects and items. However, we had adequate power to detect moderate effect sizes.

The main contribution of the paper is to demonstrate that within-subjects designs do not necessarily make research hypotheses transparent. In fact, research hypotheses are probably more opaque than we would imagine. However, there are still some research hypotheses that will be transparent to most subjects in within-subjects designs. For example, 88% of subjects were able to figure out the research hypothesis in our manipulation check. Note that this is still less than the 100% we had imagined when creating this sample experiment. Given that the risk of task transparency appears to be very low, we

do not think this should be a cause to abandon the within-subjects design. Furthermore, although there may be some risk to internal validity if the task is transparent, we argue that this risk does not outweigh the stimulus-context confound for subjective judgments in between-subjects designs. Instead, we suggest that researchers ask subjects to guess the hypothesis tested in the experiment after the study has been completed to ensure the task was not transparent. If, on the other hand, it is important that subjects understand what is being manipulated in the experiment, asking them to describe the research hypothesis can provide a type of manipulation check. The key point here is that our data indicate that neither experienced researchers nor subjects have accurate intuitions about which research hypotheses will be transparent. Therefore, we argue that this is an empirical question that can easily be tested in each experiment.

Additionally, within- and between-subjects analyses of our data produced the same conclusions. Thus, for situations in which context and stimulus are not confounded in between-subjects designs and research hypotheses are not transparent in within-subjects designs, the two types of designs may produce the same conclusions. However, because we are often unable to predict *a priori* whether stimulus-context confounds exist or research hypotheses are transparent, we argue that it important to collect data using a within-subjects design, randomizing the presentation of conditions so that both between- and within-subjects analyses can be conducted. This will provide information about the effect of design type on conclusions, which will provide a richer understanding of the effect (e.g., Frisch, 1993). For example, the use of a within-subjects design enables the researcher to ascertain subjects' perception of the normative model (e.g., Baron & Hershey, 1988). In addition, there are predictable cases in which between- and within-subjects designs result in different conclusions; see the literature on joint vs. separate evaluations for several demonstrations of this phenomenon (e.g., Hsee, Loewenstein, Blount & Bazerman, 1999).

6 Conclusions

Data from three experiments demonstrate that within-subjects designs do not regularly render the experimental task transparent. Therefore, this popular reason for rejecting the within-subjects design in favor of the between-subjects design has little empirical merit. Thus, we argue, as others have done in the past (e.g., Birnbaum, 1999), that the within-subjects design, with counterbalancing, should be the default design when measuring subjective judgments. In addition, routinely asking subjects to guess the research hypothesis upon completion of the study can allow the researcher to test for task transparency in the case that it may be important to guard against demand characteristics or provide evidence that all conditions were understood.

References

Baron, J. & Hershey, J. C. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology*, *54*, 569–579.

Bastardi, A. & Shafir, E. (1998). On the pursuit and misuse of useless information. *Journal of Personality and Social Psychology, 75,* 19–32.

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

Birnbaum, M. H. & Mellers, B. A. (1983). Bayesian inference: Combining base rates with opinions of sources who vary in credibility. *Journal of Personality and Social Psychology, 45*, 792–804.

Fischhoff, B. (1975). Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, 1, 288–299.

Fischhoff, B., Slovic, P. & Lichtenstein, S. (1979). Subjective sensitivity analysis. *Organizational Behavior and Human Performance, 23,* 339–359.

Frisch, D. (1993). Reasons for framing effects. *Organizational Behavior and Human Decision Processes, 54*, 399–429.

Grice, G. R. (1966). Dependence of empirical laws upon the source of experimental variation. *Psychological Bulletin, 66,* 488–498.

Hsee, C. K., Loewenstein, G. F., Blount, S., & Bazerman, M. H. (1999). Preference reversals between joint and separate evaluations of options: A review and theoretical analysis. *Psychological Bulletin, 12,* 576–590.

Kahneman, D. & Frederick, S. (2005). A model of heuristic judgment. *The Cambridge Handbook of Thinking and Reasoning,* New York: Cambridge University Press.

Kahneman, D., Slovic, P. & Tversky, A. (1982). *Judgment Under Uncertainty: Heuristics and Biases*. New York: Cambridge University Press.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*, 237–251.

Kirk, R. E. (1995). *Experimental Design: Procedures for the Behavioral Sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review, 72,* 407–418.

Shafir, E. (1993). Choosing versus rejecting: Why some options are both better and worse than others. *Memory & Cognition, 21*, 546–556.

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H. & Simons, A. (1991). Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social Psychology, 61,* 195–202.

Tversky, A. & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*, 453–458.

Tversky, A. & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90,* 293–315.

Tversky, A. & Kahneman, D. (1986). Rational choices and the framing of decisions. *Journal of Business*, 59*,* 251–278.

Tversky, A. & Koehler, D. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review, 101,* 547–567.

Zimmerman, D. E. (1997). A note on interpretation of the paired-samples *t* test. *Journal of Educational and Behavioural Statistics, 22,* 349–360.

Appendix A: Instructions and questions used in Experiment 1

Instructions

What follows is a brief questionnaire in which you will be asked to imagine different hypothetical scenarios and then answer a few questions about them. There are only nine questions and you should be completed in about 10 minutes. Detailed instructions are provided with each scenario. Please answer each question by drawing a checkmark in the space provided next to the option you prefer. Thank you.

Scenario 1

Imagine you serve on the jury of an only-child sole-custody case following a relatively messy divorce. The facts of the case are complicated by ambiguous economic, social, and emotional considerations, and you decide to base your decision entirely on the following few observations. To which parent would you **AWARD** sole custody of the child?

____Parent *A*	____Parent *B*
Average income	Above-average income
Average health	Very close relationship with the child
Average working hours	Extremely active social life
Reasonable rapport with the child	Lots of work-related travel
Relatively stable social life	Minor health problems

Again, imagine you serve on the jury of an only-child sole-custody case following a relatively messy divorce. The facts of the case are complicated by ambiguous economic, social, and emotional considerations, and you decide to base your decision entirely on the following few observations. Which parent would you **DENY** sole custody of the child?

____Parent *A*	____Parent *B*
Average income	Above-average income
Average health	Very close relationship with the child
Average working hours	Extremely active social life
Reasonable rapport with the child	Lots of work-related travel
Relatively stable social life	Minor health problems

Scenario 2

Imagine that the United States is preparing for the outbreak of an unusual Asian disease that is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the program are as follows:

If **Program *A*** is adopted, 200 people will be saved.

If **Program *B*** is adopted, there is a one-third probability that 600 people will be saved and a two-thirds probability that no people will be saved.

Which program do you prefer?

____Program *A*
____Program *B*

Again, imagine that the United States is preparing for the outbreak of an unusual Asian disease that is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the program are as follows:

If **Program *C*** is adopted, 400 people will certainly die.

If **Program *D*** is adopted, there is a one-third probability that no one will die and a two-thirds probability that 600 people will die.

Which program do you prefer?

____Program *C*
____Program *D*

Scenario 3

Consider the following two lotteries, described by the percentage of marbles of different colors in each box and the amount of money you win or lose depending on the color of a randomly drawn marble. Which lottery do you prefer?

Option A				
90% white	6% red	1% green	1% blue	2% yellow
$0	win $45	win $30	lose $15	lose $15
Option B				
90% white	6% red	1% green	1% blue	2% yellow
$0	win $45	win $45	lose $10	lose $15

Again, consider the following two lotteries, described by the percentage of marbles of different colors in each box and the amount of money you win or lose depending on the color of a randomly drawn marble. Which lottery do you prefer?

Option C			
90% white	6% red	1% green	3% yellow
$0	win $45	win $30	lose $15
Option D			
90% white	7% red	1% green	2% yellow
$0	win $45	lose $10	lose $15

Appendix B: Instructions and questions used in Experiment 2

Instructions

To begin, please read the following blurb:

In all experiments, the researcher has a **hypothesis**, which is what he is putting to the test. The research hypothesis is the experimenter's prediction, and the experiment is set up in a way so that the data that results will tell the experimenter whether that prediction is met or not. If the prediction is confirmed then the experimenter can say that his hypothesis is supported by the data. Let's have an example: Say your teacher decides to conduct an in-class experiment. Following a lecture on cognitive psychology she has everyone in the class take the same exam, an exam testing your knowledge of cognitive psychology. She also, however, has everyone in the class wear headphones while taking the exam. Half of the students listen to classical music, and the other half listens to top-20 favorites. In this example, the score is provided by the test everyone takes; the manipulation is whether students listen to classical or top-20 music; and the **hypothesis**, or prediction, is that students who listen to music with lyrics (top-20 music) will be more distracted and therefore do worse on the test than those who listen to classical music (music without lyrics).

In what follows there are five separate scenarios, each with their own instructions and brief set of questions. After each scenario you will be asked what exactly you think it is the experimenter is trying to learn, or in other words, what the experimenter's **hypothesis** or prediction is. You will then be asked to rate how confident you are that you are correct. And finally, you will be asked to rate each scenario as being either: 1) "Completely transparent", 2) "Somewhat transparent/somewhat opaque or 3) "Not transparent at all/completely opaque." A scenario is "transparent" if it is relatively easy to guess what the research hypothesis or prediction is. Conversely, a scenario is "opaque" if it is very difficult to figure out what is being predicted. Please answer each question by drawing a checkmark in the space provided next to the option you prefer. Thank you.

Scenario 1

Please imagine that an experimenter sits you down in a waiting room and tells you that you are to wait there until he comes and gets you. While you are waiting, a man in the waiting room says that he is trying to help his son sell raffle tickets for school. He asks you if you will purchase a $10 raffle ticket. Do you choose to:

____Purchase the $10 raffle ticket
____Decline the request

The experimenter then returns and tells you that the experiment is completed.

Again, please imagine that an experimenter sits you down in a waiting room and tells you that you are to wait there until he comes and gets you. While you are waiting, a man in the waiting room gets up and goes over to a vending machine. He comes over to you and says, "Hey excuse me. I just went to buy a pop and the vending machine accidentally gave me two. Do you want this one?" You gladly accept the can of pop. The man then says that he is trying to help his son sell raffle tickets for school.

He asks you if you will purchase a $10 raffle ticket. Do you choose to:

____Purchase the $10 raffle ticket
____Decline the request

The experimenter then returns and tells you that the experiment is completed.

We would now like you to answer the following questions about Scenario 1.

1. What do you think the researcher is trying to find out? In other words, what do think the researcher's hypothesis or prediction is for this scenario? Please write your answer in the space provided.

2. How confident are you that you have accurately guessed what the experimenter is trying to find out, what his prediction is?
   ____ 3: Extremely Confident
   ____ 2: Very Confident
   ____ 1: Somewhat Confident
   ____ 0: Neither Confident nor Unconfident
   ____−1: Somewhat Unconfident
   ____−2: Very Unconfident
   ____−3: Extremely Unconfident

3. How would you rate this within-subjects experimental scenario?
   __ 1) Completely transparent
   __ 2) Somewhat transparent/somewhat opaque
   __ 3) Not transparent at all/completely opaque

Scenario 2

Imagine you serve on the jury of an only-child sole-custody case following a relatively messy divorce. The facts of the case are complicated by ambiguous economic, social, and emotional considerations, and you decide to base your decision entirely on the following few observations. To which parent would you **AWARD** sole custody of the child?

____Parent *A*	____Parent *B*
Average income	Above-average income
Average health	Very close relationship with the child
Average working hours	Extremely active social life
Reasonable rapport with the child	Lots of work-related travel
Relatively stable social life	Minor health problems

Again, imagine you serve on the jury of an only-child sole-custody case following a relatively messy divorce. The facts of the case are complicated by ambiguous economic, social, and emotional considerations, and you decide to base your decision entirely on the following few observations. To which parent would you **DENY** sole custody of the child?

____Parent *A*	____Parent *B*
Average income	Above-average income
Average health	Very close relationship with the child
Average working hours	Extremely active social life
Reasonable rapport with the child	Lots of work-related travel
Relatively stable social life	Minor health problems

We would now like you to answer the following questions about Scenario 2.

1. What do you think the researcher is trying to find out? In other words, what do think the researcher's hypothesis or prediction is for this scenario? Please write your answer in the space provided.

2. How confident are you that you have accurately guessed what the experimenter is trying to find out, what his prediction is?
   ____ 3: Extremely Confident
   ____ 2: Very Confident
   ____ 1: Somewhat Confident
   ____ 0: Neither Confident nor Unconfident
   ____−1: Somewhat Unconfident
   ____−2: Very Unconfident
   ____−3: Extremely Unconfident

3. How would you rate this within-subjects experimental scenario?
   __ 1) Completely transparent
   __ 2) Somewhat transparent/somewhat opaque
   __ 3) Not transparent at all/completely opaque

Scenario 3

Imagine that the United States is preparing for the outbreak of an unusual Asian disease that is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the program are as follows. Which of the two programs do you favor?

____ If program A is adopted, 200 people will be saved
____ If program B is adopted, there is a one-third probability that 600 people will be saved and a two-thirds probability that no people will be saved.

Again, imagine that the United States is preparing for the outbreak of an unusual Asian disease that is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the program are as follows. Which of the two programs do you favor?

____If program C is adopted, 400 people will certainly die

____If program D is adopted, there is a one-third probability that no one will die and a two-thirds probability that 600 people will die.

We would now like you to answer the following questions about Scenario 3.

1. What do you think the researcher is trying to find out? In other words, what do think the researcher's hypothesis or prediction is for this scenario? Please write your answer in the space provided.

2. How confident are you that you have accurately guessed what the experimenter is trying to find out, what his prediction is?
   ____ 3: Extremely Confident
   ____ 2: Very Confident
   ____ 1: Somewhat Confident
   ____ 0: Neither Confident nor Unconfident
   ____−1: Somewhat Unconfident
   ____−2: Very Unconfident
   ____−3: Extremely Unconfident

3. How would you rate this within-subjects experimental scenario?
   __ 1) Completely transparent
   __ 2) Somewhat transparent/somewhat opaque
   __ 3) Not transparent at all/completely opaque

Scenario 4

Consider the following two lotteries, described by the percentage of marbles of different colors in each box and the amount of money you win or lose depending on the color of a randomly drawn marble. Which lottery do you prefer?

Option A				
90% white	6% red	1% green	1% blue	2% yellow
$0	win $45	win $30	lose $15	lose $15
Option B				
90% white	6% red	1% green	1% blue	2% yellow
$0	win $45	win $45	lose $10	lose $15

Again, consider the following two lotteries, described by the percentage of marbles of different colors in each box and the amount of money you win or lose depending on the color of a randomly drawn marble. Which lottery do you prefer?

Option C			
90% white	6% red	1% green	3% yellow
$0	win $45	win $30	lose $15
Option D			
90% white	7% red	1% green	2% yellow
$0	win $45	lose $10	lose $15

We would now like you to answer the following questions about Scenario 4.

1. What do you think the researcher is trying to find out? In other words, what do think the researcher's hypothesis or prediction is for this scenario? Please write your answer in the space provided.

2. How confident are you that you have accurately guessed what the experimenter is trying to find out, what his prediction is?
   ____ 3: Extremely Confident
   ____ 2: Very Confident
   ____ 1: Somewhat Confident
   ____ 0: Neither Confident nor Unconfident
   ____−1: Somewhat Unconfident
   ____−2: Very Unconfident
   ____−3: Extremely Unconfident

3. How would you rate this within-subjects experimental scenario?
   __ 1) Completely transparent
   __ 2) Somewhat transparent/somewhat opaque
   __ 3) Not transparent at all/completely opaque

Appendix C: Instructions and questions used in Experiment 3

Instructions

Hello and thank you for participating. In this questionnaire you will be presented with five brief scenarios from SJDM literature. You will be shown the questions for each scenario as they would appear to research subjects in a within-subjects design, in which the relevant comparisons to be made are made by the same research subjects at the same time. The purpose of this brief questionnaire is to assess which of the following experimental scenarios you, as a researcher, estimate would be perceived as "transparent" if presented within subjects to a sample of typical psychology research subjects. After answering all of the questions for each within-subjects scenario, you will be asked to estimate whether you think

the scenario would be either "completely transparent", "somewhat transparent/somewhat opaque" or "not transparent at all/completely opaque" to one of your typical research subjects. Here we will assume an experiment is "transparent" if it would seem to be relatively easy for your typical subject to "see through" the research design by guessing what the research hypothesis or prediction is. Conversely, we will assume a design is "opaque" if it seems to be very difficult for your typical subject to figure out what the researcher is up to by guessing what is being predicted. When making your rating please keep in mind that each scenario presented below is to be considered within subjects. Thank you for your time.

Scenario 1

Imagine you serve on the jury of an only-child sole-custody case following a relatively messy divorce. The facts of the case are complicated by ambiguous economic, social, and emotional considerations, and you decide to base your decision entirely on the following few observations. To which parent would you ***award*** sole custody of the child?

____Parent *A*	____Parent *B*
Average income	Above-average income
Average health	Very close relationship with the child
Average working hours	Extremely active social life
Reasonable rapport with the child	Lots of work-related travel
Relatively stable social life	Minor health problems

Again, imagine you serve on the jury of an only-child sole-custody case following a relatively messy divorce. The facts of the case are complicated by ambiguous economic, social, and emotional considerations, and you decide to base your decision entirely on the following few observations. To which parent would you ***deny*** sole custody of the child?

____Parent *A*	____Parent *B*
Average income	Above-average income
Average health	Very close relationship with the child
Average working hours	Extremely active social life
Reasonable rapport with the child	Lots of work-related travel
Relatively stable social life	Minor health problems

How would you rate this within-subjects experimental scenario?

__ 1) Completely transparent
__ 2) Somewhat transparent/somewhat opaque
__ 3) Not transparent at all/completely opaque

Scenario 2

Imagine that the United States is preparing for the outbreak of an unusual Asian disease that is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the program are as follows. Which of the two programs do you favor?

____If program A is adopted, 200 people will be saved
____If program B is adopted, there is a one-third probability that 600 people will be saved and a two-thirds probability that no people will be saved.

Again, imagine that the United States is preparing for the outbreak of an unusual Asian disease that is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the program are as follows. Which of the two programs do you favor?

____If program C is adopted, 400 people will certainly die
____If program D is adopted, there is a one-third probability that no one will die and a two-thirds probability that 600 people will die.

How would you rate this within-subjects experimental scenario?
__ 1) Completely transparent
__ 2) Somewhat transparent/somewhat opaque
__ 3) Not transparent at all/completely opaque

Scenario 3

Consider the following two lotteries, described by the percentage of marbles of different colors in each box and the amount of money you win or lose depending on the color of a randomly drawn marble. Which lottery do you prefer?

Option A				
90% white	6% red	1% green	1% blue	2% yellow
$0	win $45	win $30	lose $15	lose $15
Option B				
90% white	6% red	1% green	1% blue	2% yellow
$0	win $45	win $45	lose $10	lose $15

Again, consider the following two lotteries, described by the percentage of marbles of different colors in each box and the amount of money you win or lose depending

on the color of a randomly drawn marble. Which lottery do you prefer?

Option C			
90% white	6% red	1% green	3% yellow
$0	win $45	win $30	lose $15
Option D			
90% white	7% red	1% green	2% yellow
$0	win $45	lose $10	lose $15

How would you rate this within-subjects experimental scenario?
__ 1) Completely transparent
__ 2) Somewhat transparent/somewhat opaque
__ 3) Not transparent at all/completely opaque

How would you rate this within-subjects experimental scenario?
__ 1) Completely transparent
__ 2) Somewhat transparent/somewhat opaque
__ 3) Not transparent at all/completely opaque