ARE WITHIN-SUBJECTS DESIGNS TRANSPARENT?

A Dissertation by

Charles Garrett Lambdin

M.A., Wichita State University, 2006

B.A., Wichita State University, 2003

Submitted to the Department of Psychology
and the faculty of the Graduate School of
Wichita State University
in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

May 2009

© Copyright 2009 by Charles Garrett Lambdin

All Rights Reserved

ARE WITHIN-SUBJECTS DESIGNS TRANSPARENT?

The following faculty members have examined the final copy of this dissertation for form and content, and recommend that it be accepted in partial fulfillment of the requirement for the degree of Doctor of Philosophy with a major in Psychology.

Victoria Shaffer, Committee Chair

Alex Chaparro, Committee Member

Edgar Merkle, Committee Member

Evan Palmer, Committee Member

Robert Feleppa, Committee Member

Accepted for the College of Liberal Arts and Sciences

William D. Bischoff, Dean

Accepted for the Graduate School

J. David McDonald
Associate Provost for Research
and Dean of the Graduate School

DEDICATION

To my beloved wife, Candace Gilmore

ABSTRACT

In the field of judgment and decision making (JDM), it is sometimes argued that within-subjects designs should be avoided because they are transparent. This claim ignores that between- and within-subjects designs are frequently not interchangeable. The transparency of within-subjects replications of four between-subjects JDM experiments was empirically tested. None of the replications were found to be transparent, and further, transparency did not in any way affect the success or failure of the replications. It was also found that members of the Society for Judgment and Decision Making were not able to predict how transparent tasks would be when presented within subjects, suggesting that researchers have no special insight into what will or will not prove transparent to participants.

TABLE OF CONTENTS

Chapter	Page
I. INTRODUCTION	1
Two assumptions in JDM literature	2
Assumption one: Within-subjects designs are transparent	3
Assumption two: Between- and within-subjects studies are interchangeable	5
Design type as an independent variable	6
Between-subjects context effects	9
Between-subjects context effects in published studies	13
Between- and within-subjects designs can be very different animals	15
Transparency and between-subjects context effects aside...	17
Within- and between-subjects hypotheses	19
Are within-subjects designs a continuum of transparency?	21
Between- vs. within-subjects designs: Other issues and controversies	23
Introduction conclusions	27
II. Experiment 1	31
Method	33
Results	36
Discussion	40
III. Experiment 2	44
Method	48
Results	51
Discussion	68
IV. Experiment 3	72
Method	73
Results	77
Discussion	105
V. Experiment 4	112
Method	113
Results	114
Discussion	124
VI. General Discussion	129

TABLE OF CONTENTS (continued)

Chapter	Page
Research summary	129
Implications for and discussion of common objections to the within-subjects design	133
Conclusion	136
VII. References	142
VIII. Appendices	148
Appendix A	149
Appendix B	154
Appendix C	164

LIST OF TABLES

Table	Page
1. Average Confidence Ratings, DNST vs. ST	81
2. Percent Identified Prediction, Different vs. Same Page	86
3. Average Confidence Ratings, Different vs. Same Page	90
4. Percent Identified Prediction	91
5. Average Confidence Ratings	97
6. SJDM Member Transparency Ratings	119

LIST OF FIGURES

Figure	Page
1. Within-subjects replication data, the child custody case	37
2. Within-subjects replication data, the Asian disease problem	39
3. The marbles lotteries, first vs. second option presented as dominating lottery	52
4. The child custody case, first vs. second parent presented as “enriched option”	53
5. The Asian disease problem, first vs. second program presented as gamble	55
6. The marbles lotteries, different vs. same page	56
7. The child custody case, different vs. same page	57
8. The Asian disease problem, different vs. same page	58
9. Within-subjects replication data, the marbles lotteries	60
10. Within-subjects replication data, the child custody case	63
11. Within-subjects replication data, the Asian disease problem	64
12. The marbles lotteries, within vs. between subjects	66
13. The child custody case, within vs. between subjects	67
14. The Asian disease problem, within vs. between subjects	68
15. The manipulation check transparency ratings, DNST vs. ST	78
16. The marbles lotteries transparency ratings, DNST vs. ST	79
17. The Princeton applicant transparency ratings, DNST vs. ST	79
18. The child custody case transparency ratings, DNST vs. ST	80
19. The Asian disease problem transparency ratings, DNST vs. ST	81
20. The marbles lotteries data, DNST vs. ST	82
21. The child custody case data, DNST vs. ST	84
22. The Asian disease problem data, DNST vs. ST	85
23. The manipulation check, transparency ratings for different vs. same page	87
24. The marbles lotteries, transparency ratings for different vs. same page	88
25. The Princeton applicant, transparency ratings for different vs. same page	88
26. The child custody case, transparency ratings for different vs. same page	89
27. The Asian disease problem, transparency ratings for different vs. same page	90
28. The manipulation check transparency ratings	93
29. The marbles lotteries transparency ratings	94
30. The Princeton applicant transparency ratings	95
31. The child custody case transparency ratings	96
32. The Asian disease problem transparency ratings	96
33. The marbles lotteries, Experiment 2 vs. Experiment 3	99
34. The child custody case, Experiment 2 vs. Experiment 3	100
35. The Asian disease problem, Experiment 2 vs. Experiment 3	101
36. The marbles lotteries data, Experiments 2 & 3 combined	102
37. The child custody case data, Experiments 2 & 3 combined	104
38. The Asian disease problem data, Experiments 2 & 3 combined	105
39. The marbles lotteries data, undergraduates vs. SJDM members	115
40. The child custody case data, undergraduates vs. SJDM members	116
41. The Asian disease problem data, undergraduates vs. SJDM members	118

LIST OF FIGURES (continued)

Figure		Page
42.	The marbles lotteries transparency ratings, undergraduates vs. SJDM members	121
43.	The Princeton applicant transparency ratings, undergraduates vs. SJDM members	122
44.	The child custody case transparency ratings, undergraduates vs. SJDM members	123
45.	The Asian disease problem transparency ratings, undergraduates vs. SJDM members	124

CHAPTER 1

INTRODUCTION

Forty-three years ago, Grice (1966) observed that between- and within-subjects designs used to investigate the same phenomenon commonly do not produce the same findings. Greenwald (1976) added that, despite this fact, researchers very rarely base their choice of design type upon its appropriateness to the phenomenon investigated. In the field of Judgment and Decision Making (JDM), a controversy has long existed over the merits and demerits of between- and within-subjects designs. Kahneman and Frederick recently stated that this debate is a “highly contentious issue” (2005, p. 280).

Birnbaum, for example, has argued that in the standard JDM experiment, the between-subjects design is wholly inappropriate (1982, 1999). Parducci (1965; 1971) argued that when participants are asked to rate a stimulus on a scale (i.e., “small,” “medium,” “large,” etc.), that their subjective ratings are the result of two factors: 1) the relative position of the stimulus to the perceived endpoints of its scale and 2) the stimulus’ percentile rank. The former is the range and the latter is the frequency, hence the name: Paducci’s range-frequency model. If the frequency is the same for two stimuli but their perceived positions on their respective scales (or contexts) are different, they will be assigned different ratings by participants. Thus, if in a between-subjects design the perceived scale endpoints for two stimuli are very different between groups, the two stimuli may be assigned highly different ratings. Birnbaum’s addition to Parducci’s (1965; 1971) theory is that all stimuli will be perceived to exist within a context, whether that context—or range—is specified or not. If the context or range is not made specific, then each stimulus will likely evoke its own context. If two stimuli evoke different contexts between groups, then those groups are not comparable. This poses a major impediment to use of some between-

subjects designs. We should not be surprised to find, between subjects, that the number 450 was rated as being larger than the number 550 if the contexts evoked in each group differ (e.g., Varey, Mellers & Birnbaum, 1990).

Thus, Birnbaum adopted Parducci's (1965) range-frequency model to argue that between-subjects designs typically do not produce meaningful comparisons when the dependent variable is a subjective judgment. Kahneman and Frederick, however, have argued that "between-subjects designs are more appropriate for the study of heuristics and judgment" (2005, p. 280). One of the main reasons they gave for this claim is that "factorial designs are transparent" (p. 280). Furthermore, the assumption that within-subjects designs are transparent seems to be somewhat widely held in the field of JDM (Bastardi & Shafir, 1998; Fischhoff, Slovic & Lichtenstein, 1979; Kahneman & Frederick, 2005; Tversky & Kahneman, 1983). What, then, does it mean exactly to say that a within-subjects design is transparent? In the between-subjects design, different participants are used in each condition of an experiment. Since a single participant sees only one condition and no other, the between-subjects design effectively leaves research participants in the dark regarding the comparisons the experimenter wishes to make. In the within-subjects design, however, the same participants are utilized in every condition. Therefore, the experimenter reveals the comparisons being made to every participant. If participants then have an easier time discerning what the research hypothesis is, it can be said that the within-subjects design is more "transparent" than the between. This would be a concern to researchers given that if most participants know what the researcher is looking for, the resulting demand characteristics could confound the results of the experiment, creating low internal validity.

Two assumptions in JDM literature

There are two large assumptions regarding between- and within-subjects designs that can be found in JDM literature (though they are certainly not limited to the field of JDM). The first is that while the between-subjects design is not transparent, the within-subjects design is (e.g., Fischhoff, Slovic & Lichtenstein, 1979; Tversky & Kahneman, 1983), and should be avoided because of this due to resulting confounds (Kahneman & Frederick, 2005). The second assumption is that—issues of transparency aside—between- and within-subjects designs are exchangeable; i.e., they can interchangeably be used to experimentally test the same research hypotheses (e.g., Bastardi & Shafir, 1998; Fischhoff, Slovic & Lichtenstein, 1979; Kuhberger, Komunska & Perner, 2001).

Though it is easily demonstrable that the second of these assumptions is false, this does not seem to be a widely appreciated point (Greenwald, 1976; Grice, 1966). Furthermore, it is easily seen that these assumptions conflict with each other. Anyone making both of them is in a precarious position. The purpose of this paper is to put the first assumption to empirical test and to remind readers that the second assumption is plainly false.

Assumption one: Within-subjects designs are transparent

The assumption that within-subjects designs are transparent will be addressed by four experiments designed to test for the transparency of experimental tasks presented within subjects. An experiment is considered “transparent” when it is relatively easy for participants to see through the research design (Tversky & Kahneman, 1983). In this set of experiments, transparency will be operationalized as the ability of participants to effectively guess the research hypothesis, or in other words, what is being predicted. One could argue that in order for a design to be considered transparent, participants must guess both what is being predicted as well as the direction of that prediction. Here, to be as liberal as possible, we will assume that merely

figuring out the research hypothesis results in transparency. Some (e.g., Kahneman & Frederick, 2005) have claimed that in a within-subjects design, participants automatically know what the manipulation is and that in JDM experiments this is enough to confound the study. Evidence presented in this paper casts serious doubt onto this claim.

It should be added that the contention here is not that within-subjects designs are not more transparent than between. That is not what is being argued. Obviously, many within-subjects designs would have to be somewhat more transparent than between-subjects designs. The contention, rather, is that this transparency is often greatly exaggerated and is typically not of a magnitude that would actually alter the results of a study.

Addressing the assumption that within-subjects designs are transparent is important for two reasons. First, it can be assumed that this assumption is somewhat prevalent in the field of JDM. One of its top researchers (Daniel Kahneman) explicitly stated that heuristics and judgments should be tested between subjects primarily because of the transparency of within-subjects designs (Kahneman & Frederick, 2005). A major problem with this statement involves the assumption that between- and within-subjects designs (transparency issues aside) are interchangeable. This assumption is false since between- and within-subjects designs typically produce different results and the appropriateness of each design is highly context dependent. Ignoring this point and simply sticking with one design over the other as a matter of habit would produce many invalid studies and thereby have a deleterious impact on the literature as a whole.

If a JDM researcher uses a between-subjects design because of the assumption that within-subjects designs are transparent, or because of the assumption that between-subjects designs have greater ecological validity, then s/he is ignoring the fact that the choice of design type must ultimately be dependent upon the nature of the phenomenon investigated.

Furthermore, because of the assumption of transparency, some researchers may be opting for the more costly, more cumbersome, less powerful and sometimes less valid between-subjects design for no good reason.

The second reason that it is important to address the assumption that within-subjects designs are transparent is that it does not seem that this alleged transparency has ever been put to a direct empirical test. Though many authors comment on this assumed transparency, this author believes that this issue is often significantly overstated. This belief is based on many experiences replicating between-subjects experiments using within-subjects designs. The assumption that within-subjects designs are transparent will be directly empirically tested in a series of four experiments. The second assumption, that issues of transparency aside, between- and within-subjects designs are interchangeable, will be addressed now.

Assumption two: Between- and within-subjects studies are interchangeable

It is likely that many researchers would argue against the assumption that between- and within-subjects designs are interchangeable simply because of the assumed transparency of the latter. The two designs, however, are often not interchangeable for other reasons. When one takes transparency out of the equation, between- and within-subjects designs still frequently do not test the same hypotheses. If these designs are often not interchangeable, it follows that the conditions for the appropriate use of each will often differ. It further follows that—to the extent to which this fact is left unappreciated—there will be instances of the inappropriate use of each design. Therefore, to sweepingly conclude, as Kahneman and Frederick (2005) have, that between-subjects designs should be preferred in general is irresponsible and inaccurate.

The assumption that the use of alternative experimental designs to investigate the same phenomenon will naturally yield corroborating results wholly ignores the fact that the results of

any study are in part a function of the experimental design employed. An experimental finding is a function of what is being looked at as well as how we choose to look at it. Ignoring the latter side of this equation blinds us to the “methods-dependent nature” of any experimental finding. Exclusively employing only the between-subjects design would commit us to what Cook and Campbell (1979) called the “mono-method bias.” Replication using different methods helps us to avoid this (Sidman, 1960). (And as some have pointed out, many of psychology’s famous findings—when we bother to look—actually do not replicate [Lykken, 1991].) If a particular finding appears only with the use of a particular design, understanding why this is the case will help us explicate the conditions necessary for the effect in question to exist, or will help us determine whether the effect is even real, especially if it turns out that one of the designs was an inappropriate choice.

Design type as an independent variable

In psychology, it is commonly the case that results disappear altogether or even reverse themselves when switching from a between- to a within-subjects design (Birnbaum, 1982). Because of this, some have argued that design type should often be treated as an explicit independent variable (e.g., Fischhoff, Slovic & Lichtenstein, 1979; Keren, 1993). If a finding is particular to a specific design, in order to meaningfully claim it is empirically supported it is crucial to flesh out exactly why alternate designs do not produce corroborating results. This stresses the value of replication. It may be pertinent to add that replications are important and informative regardless of whether or not they are successful, even if peer-review journals continue to fail to appreciate this fact (Mahoney, 1977).¹ A potential problem with this argument

¹ Some interesting (wholly ignored) work has been done on this topic. It implies that the entire peer-review process may be far less scientific than many assume (if scientific at all). Even though results

is that though the treatment of design type as an explicit independent variable is certainly an interesting idea, seeing if effects hold both between- and within-subjects can only validly be done if both designs are appropriate tests of the phenomenon in question.

This can be seen by looking at the controversy over the concept of “base-rate neglect.” The idea behind the base-rate neglect fallacy is that people routinely ignore base-rate information when it is pertinent to the decision at hand. For example, if a patient has three out of five symptoms consistent with illness *X* and four out of five symptoms consistent with illness *Y*, the idea is that most doctors will diagnose the patient as having illness *Y*, even if the base rate of *Y* in the population is only 2% and the base rate of *X* is 20%. The problem is that the base-rate neglect fallacy that is found in between-subjects research does not replicate in within-subjects designs, in which participants make good use of base-rate information (Birnbaum & Mellers, 1983; Varey, Mellers & Birnbaum, 1990).

There are different interpretations of why this is the case. Fischhoff, Slovic and Lichtenstein (1979) found, for example, that when replicating two standard base-rate neglect problems within subjects, 65% of participants ordered their probability assessments according to

should be irrelevant, reviewers tend to only approve papers for publication that are consistent with their own biases (Mahoney, 1977). Additionally, who researchers choose to cite seems to be largely influenced by two things: 1) collusion or “mutual back-scratching,” meaning that who an author cites is influenced by who is willing to cite them in return (Sorokin, 1956) and 2) “monkey see monkey do,” meaning that most researchers simply cite papers they see other researchers citing (Taleb, 2007). This has three disturbing implications that together cast much doubt on the entire peer-review process: 1) most researchers do not actually read the papers they cite, 2) the majority of peer-review publications are never read by anyone, and this implies 3) whether any given paper becomes well cited or not is largely due to chance and not scientific merit or veracity (Taleb, 2007). If a paper receives an early advantage by being read and cited, other authors will see it cited and then also cite it themselves. The number of citations is then given to exponential increase. As most papers are not read at all, the papers that become “seminal” become so largely due to chance. This and other findings have led some to make the strong claim (e.g., Andreski, 1972) that the ascendancy of the “publish or perish” reality in academia makes many academics conduct much unscientific research simply to professionally survive.

the base rates, a far cry from the claim that people routinely wholly ignore base-rate information. They speculate that a single piece of diagnostic information (as seen in a between-subjects design) often merely serves as an anchor, whereas seeing a range of data (as seen in a within-subjects design) aids in breaking one away from that anchor. Thus, it has been argued, looking at this problem both between and within subjects helped us understand its causes as well as ways to lessen its impact (Fischhoff, Slovic & Lichtenstein, 1979). Further, if the effect was only tested for between subjects, the actual narrow specificity of the conditions necessary to produce it would remain masked due to the confines of the design relied on.

Similarly, replicating hindsight-bias studies within subjects—and thereby exposing participants to alternative values—serves to reduce the effect of the bias (Slovic & Fischhoff, 1977). The hindsight bias describes our inability to disregard knowledge once learned, our inability to return to a former state of ignorance (Fischhoff, 1975). The hindsight bias, for example, greatly impacts the verdicts of jury members, who make their decisions from the vantage point of hindsight (Hawkins & Hastie, 1990). Fischhoff, Slovic and Lichtenstein (1979) found that replicating an experiment on sensitivity to sample size information within subjects did not, however, change the results at all. They argued that this suggests that this particular bias is far more robust than either hindsight or base-rate neglect.

Fischhoff, Slovic and Lichtenstein (1979) claimed that in this way, replicating a between-subjects result within subjects acts as a “subjective sensitivity analysis.” The argument is that a successful within-subjects replication of a between-subjects finding is a demonstration of the robustness of the effect. This argument, however, is dependent upon the extent to which the within-subjects replication is in fact more transparent than the between-subjects original. If the level of transparency of the within-subjects replication is not of a magnitude that would actually

impact the overall pattern of results, then it does not serve as a subjective sensitivity analysis. It would simply be a replication. Fischhoff, Slovic and Lichtenstein's (1979) argument also assumes that the between- and within-subjects variations are both appropriate designs and that they test the same hypotheses.

Between-subjects context effects

If between- and within-subjects designs often do not test the same hypotheses for reasons other than task transparency, what are these other reasons? One of them is the between-subjects context effect. Parducci (1965; 1971) demonstrated that manipulating the context a stimulus appears in results in the same stimulus being assigned different subjective values by different groups. Birnbaum (1982, 1999) has argued that Parducci's (1965) range-frequency model demonstrates that between-subjects designs do not allow a valid comparison of subjective judgments whenever different contexts exist in different groups of participants. If the same exact stimulus is given different subjective values by different groups, then a between-group comparison of different stimuli is uninterpretable and therefore meaningless (Birnbaum, 1999). This is because if unspecified contexts are allowed to exist between groups, then stimulus and context become confounded. When stimulus and context are confounded and stimuli evoke their own contexts, this often generates paradoxical conclusions (Varey, Mellers & Birnbaum, 1990). Such paradoxes are likely what led Hertwig to call between-subjects designs "strange animals" (Mellers, Hertwig & Kahneman, 2001, p. 273). Hereinafter, the confounding of stimulus and context will be referred to as a "between-subjects context effect."

To illustrate, suppose one wants to use participants' responses to measure the subjective psychological value of certain stimuli using two groups in a between-subjects design. A point that is little appreciated is that if different contexts exist in the between-subjects groups, the

averaged responses of each group will not provide an ordinal scale of subjective value (Birnbaum, 1999). Thus, a between-subjects design would not even allow for a comparison of the stimuli. If this point is ignored, the discussion of the result will draw an inappropriate conclusion from an invalid comparison. This would be analogous to comparing two numbers while being wholly oblivious to the fact that they are from different scales. Birnbaum (1999) demonstrated this by conducting a between-subjects experiment in which two groups were asked to rate how large a number was on a 10-point scale, with one = “very very small” and 10 = “very very large.” One group was asked to rate how large the number nine was and the other the number 221.

The average rating for the number nine was 5.13 and the average for 221 was 3.10, $t(83) = 3.52, p < .001$. Thus, participants significantly rated the number nine as being larger than the number 221 (Birnbaum, 1999). Why? Because, Birnbaum suggests, in the “nine” group participants likely made their ratings from within a single-digit context, whereas in the “221” group there was a triple-digit context. Thus, because of this context effect, these between-subjects results cannot be meaningfully compared. Varey, Mellers and Birnbaum (1990), in a similar experiment, found that participants rated the number 450 as being larger than 550. There was obviously not a single-digit context in the “450” group, but a context effect must still have been present—unless we would rather conclude that participants really thought 450 was larger than 550. The point is that between-subjects context effects are often present and that between-subjects designs often generate “wrong” results, even when unrecognized.

Birnbaum chose numbers as stimuli for two very important reasons. First, he stated that he wanted to use a very obvious example to show that between-subjects comparisons of judgments produce “wrong” conclusions so that more people might object when judgments are

compared between subjects in less obvious and more complex settings. Second, he chose numbers so that the usual explanations for between- vs. within-subjects differences in results could not be employed (1999). All too often, the odd differences in between- and within-subjects results are hastily explained away by faulting the within-subjects design. Social psychologists, for example, often argue that within-subjects designs are transparent and therefore should be avoided due to the threat of demand characteristics (Birnbaum, 1999). Because of this, between-subjects designs are heavily relied upon and “It may be that many of the counterintuitive findings so well-liked in social psychology are merely results of the *confounding of contexts* that can occur in between-subjects designs” (Birnbaum, 1982, p. 442, italics in the original). Another common argument is that because of their level of disclosure, within-subjects designs lack ecological validity (Fischhoff, Slovic & Lichtenstein, 1979; Kahneman & Frederick, 2005). Birnbaum used numbers so that such common arguments could not be summoned. His concern is that in instances where these popular arguments seem to “fit,” where they produce an explanatory narrative that sounds good on the surface and seems plausible, they are adopted in lieu of the more commonly correct explanation: between-subjects context effects.

To illustrate Birnbaum’s (1999) point: Between subjects, nine is rated as being larger than 221. Within subjects, 221 is rated as being larger than nine. The standard explanation for this difference would run as follows: “Within subjects, the task is made transparent and so people do not respond intuitively, as they normally would. When presented within subjects, a demand characteristic was created that led people to switch their answers. The between-subjects design has greater ecological validity as it captures that, when left to their casual, everyday modes of thinking, most people do in fact feel that nine is larger than 221.” Birnbaum’s examples illustrate that such odd (indeed wholly nonsensical) results are only made explicable—

and in fact become expected—when one theorizes the existence of different contexts in between-subjects designs (Birnbaum, 1999).

Birnbaum's (1999) conclusion was that all studies with a dependent variable of human judgment should evoke our skepticism if conducted between subjects. Though this argument does not apply to between-subjects studies using objective dependent variables (standardized test scores, days with fever, etc.), if the dependent variable involves human judgment of any kind, context effects will likely confound a between-subjects comparison—even if the study is double blinded. For instance, assume that some participants are placed in a placebo group and some in a treatment group in a double-blind study and are asked to assess the treatment received. They obviously cannot, since there is no comparison. They are all stuck within their own contexts. And yet the “study hopes to learn something new by combining judgments from many people, each of whom cannot, in principle, answer the question” (1999, p. 248). Because of this, if a between-subjects result regarding judgment does not replicate within subjects, there may very well be an empirical reason to be skeptical about the claim being made (Birnbaum & Mellers, 1983). Birnbaum (1999) wrote that though multiple studies have (he feels) conclusively made this point with a wide variety of stimuli and judgment scales, applied researchers commonly ignore it. He amusingly referenced a between-subjects questionnaire that found that farm practices were increasing the grain supply in the Soviet Union while the actual grain supply was decreasing (Birnbaum, 1999; Medvedev, 1971).

Though the author thinks that Birnbaum may be overstating the problem by sweepingly claiming that between-subjects studies should never be used to investigate judgment and heuristics, the point remains that between-subjects context effects are a serious and severely underappreciated concern. Consider: Within subjects, it is easy to deal with context effects

(such as order effects) by simply counterbalancing. Between-subjects context effects, however, are very difficult to control for and in some experiments will probably not even be recognized even when present. They therefore pose a much more serious and at the same time less appreciated threat.

Between-subjects context effects in published studies

The above examples are certainly interesting, but they nevertheless leave it somewhat difficult to imagine how this problem might plague the kinds of studies that are more commonly conducted. Therefore, it is illustrative to take a closer look at the impact of between-subjects context effects in a few real examples. A famous example is Jones and Aronson's (1973) famous just world hypothesis finding. The finding is that respectable women are judged to be at greater fault of causing their own rape than less respectable women.² This finding can be reversed by switching from a within- to a between-subjects design (Birnbaum, 1982). Birnbaum argued this reversal occurs because in the between-subjects design the stimulus is confounded with the context. For instance, in the Jones and Aronson (1973) within-subjects design, participants are comparing virgins to divorcees; whereas in the between-subjects design, participants in the "virgin" condition are not given the relevant comparison and are likely comparing the virgin in question to other hypothetical virgins rather than divorcees (Birnbaum, 1982).

To take another example, Bastardi and Shafir (1998), after describing a series of between-subjects studies, admit to readers that much of what they found is largely attributable to

² This is because, presumably, less-respectable women are perceived as "deserving" it more, which means that a respectable woman must have done something to lead the rapist on. Incidentally, I highly doubt this finding would replicate today.

context effects. Bastardi and Shafir asked participants to imagine they were on the admissions committee of Princeton University. They asked participants in one group whether they would accept or reject an applicant who plays varsity soccer, has good letters of recommendation, is editor of the school paper, has a combined SAT score of 1250 and a high school average of B. Fifty-seven percent of people accepted the applicant. They gave another group of participants the same scenario, excepting that here there were conflicting reports regarding the high school average grade, one stating B and the other A. Participants could accept or reject the applicant, or wait for the grade discrepancy to be cleared up. Seventy-four percent of participants chose to wait. When they were then told that the high school average is B, 49% rejected the applicant and only 25% accepted.

After noting these results (and a similar, more dramatic result in another variation) they asked readers to note that this effect is largely due to a “contrast” effect (p. 23, 1998). Namely, when participants are told that the applicant plays varsity soccer, is editor of the paper and has a B average, the B average seems more impressive than it does to those who are in the other group, where there is a possibility that the average grade is actually A. What is particularly fascinating about this example is that what is almost explicitly being tested for is the presence of a between-subjects context effect. Thus, admitting its presence does not damage the finding. Normally the presence of such an effect is best left ignored, as acknowledging its presence would call the success of the study into doubt.

Bastardi and Shafir (1998) stated that they performed a within-subjects replication of such a study to see if a preference reversal would still occur within subjects, since showing the task within subjects supposedly makes it transparent. They found that 56% of the sample chose to wait for noninstrumental information and that only 3% (one participant) switched preferences.

What is of interest here is that Bastardi and Shafir, if I may respectfully suggest so, seem to fail to appreciate that the within-subjects replication does not test the same hypothesis as the between-subjects original. The between-subjects experiment tests for a between-groups preference reversal caused by a between-subjects context effect. The within-subjects experiment tests for whether participants who already accepted the applicant will then wait for what should now be obviously irrelevant information. This is a test for disjunction effects.

Between- and within-subjects designs can be very different animals

If a study is conducted between subjects and the stimuli in each group evoke their own contexts, then the participants in each group are trapped within their respective contexts. The stimuli cannot be meaningfully compared without letting the participants directly compare the two stimuli themselves.³ There is almost an irony in the between-subjects context effect; it produces a situation in which participants are asked to judge something without being told the context for the judgment. Birnbaum (1982, p. 444) thus compares the between-subjects design to an old stand up routine:

Person 1: “How’s your wife?”

Person 2: “Compared to what?”

This is much more than an amusing anecdote; it is a particularly illuminating example. Imagine that you are doing an experiment in which you want to know how well husbands get along with their wives compared to how well they get along with their mothers-in-law as well as

³ This is not entirely accurate. There are elaborate ways to guard against between-subjects context effects, such as designing a triple-blind study (Birnbaum, 1999). (Here, the experimenter does not know what group participants are in, participants do not know what group they are in and independent judgment raters do not know what group the participants are in.) It is certainly easier, however, to simply conduct the study within subjects.

their own mothers. Imagine you do this experiment within subjects. You ask husbands to rate how well they get along with their wives, and, you also ask them to rate how well they get along with their mothers-in-law and then their mothers.

Now imagine you do this experiment between subjects. The results may be very different. In fact, why on earth would you expect them to be similar? Here you ask one group of husbands to rate how well they get along with their wives. Then you ask a second group of husbands to rate how well they get along with their mothers-in-law. Then you ask a third group of husbands to rate how well they get along with their mothers. Since they do not see all of the relevant questions, many of them are likely rating how well they get along with their mothers-in-law compared to some idiosyncratic notion of “the average mother-in-law,” to how well other husbands are imagined to get along with their mothers-in-law. This is entirely different from rating how well they get along with their mothers-in-law compared to how well they get along with other people they know. Because of this, and contrary to what is often assumed, the within- and between-subjects assessment may not even address the same questions. This has a broad implication regarding the appropriateness of each design.

This also brings up an important point: One of the most vital steps in writing up one’s results is the interpretation of what they mean. This is always a qualitative endeavor; it cannot be quantified. It requires critical thinking and sometimes a number of assumptions. One of these assumptions is what exactly was going on in participants’ minds. As the example above points out, how participants think about a problem can be very different in between- and within-subjects designs. This not only impacts the appropriateness of each design but also the ways in which the results they yield can be meaningfully interpreted. Birnbaum (1999), for example, points out that student evaluations of teachers are done between subjects; this is the only realistic

way in which they can be done. But, since they are performed between subjects, students are probably comparing how well their teacher taught the class they just took to how well other teachers they've had taught other classes they've taken. This is not actually all that meaningful. What would be meaningful is a within-subjects design, which would result in students comparing how well their teacher taught the class they just took to how well other teachers would do teaching the exact same class. Though such a design is clearly not feasible, this example does point out that in between- and within-subjects designs, the comparisons that participants end up making are often very different.

Transparency and between-subjects context effects aside...

In discussing the assumption that—transparency aside—between- and within-subjects designs are interchangeable, we have seen that this statement is sometimes false because of between-subjects context effects. We will now discuss other reasons why between- and within-subjects studies might not be interchangeable. Moving beyond transparency and between-subjects context effects, within- and between-subjects designs sometimes test different hypotheses (and will therefore be appropriate or inappropriate to use) due to the vagaries of how the hypothetical construct in question is operationally defined.

Take, for example, research on the disjunction effect. A “disjunction effect” is defined as a violation of Savage’s (1954) “sure thing principle” (STP), which states that if one prefers doing x given A occurs and one prefers doing x given A does not occur, then one should prefer doing x even when the outcome of A is unknown (or unknowable). Using mostly within-subjects designs, Tversky and Shafir (1992) found that participants tended to violate Savage’s STP in two-step gambles. Kuhberger, Komunska and Perner (2001), using mostly between-subjects

designs, found that participants do not violate the STP and concluded there is no evidence that the disjunction effect exists.

One of the problems with Kuhberger, Komunska and Perner's (2001) claim is that a between-subjects design does not allow a person to violate the STP. Since a disjunction effect is simply a STP violation, between-subjects data are irrelevant. In order for a participant to violate the STP in a two-step gamble the following three events must occur: 1) s/he must accept a second bet after having won the first, 2) accept it after having lost the first and 3) then not accept it when the outcome of the first bet is unknown. It does not make sense to derive this pattern between subjects and then call the individual responses of three different people a "STP violation." In a between-subjects design we do not actually know if these three people—had they been exposed to all three necessary conditions—would have actually responded in a way consistent with a STP violation. Thus, what may look like a single STP violation between subjects might actually be three adherences to the STP within subjects (Lambdin & Burdsal, 2007). Lambdin and Burdsal (2007) examined the disjunction effect in two-step gambles within subjects, presenting all three conditions to the same participants and found that when participants' responses are analyzed individually it was in fact the case that the STP was regularly violated, providing direct support for the existence of the disjunction effect.⁴ Another example of a problematic use of the between-subjects design is Shafir (1993), in which, to take but one example, participants were asked to pretend they were jurors in a child custody case.

⁴ Bagassi and Macchi (2006) came up with an original way to look at disjunction effect; unfortunately, they inappropriately used a between-subjects design. This error was corrected by Sun, Li and Li (2008), who replicated Bagassi and Macchi's (2006) work using a within-subjects design. Interestingly, though Sun, Li and Li (2008) agreed with Lambdin and Burdsal (2007) that disjunction effects can only validly be tested for within subjects, they assumed that doing so renders the task transparent.

They were presented with a brief description of two parents, “Parent A” and “Parent B.” Half of the participants were asked which parent they would award custody to and the other half were asked which parent they would deny custody to. Shafir hypothesized that since Parent B had both more extreme positive and negative traits, that Parent B constituted the “enriched” option. Thus, Parent B—according to the theory of reason-based choice—would likely be both awarded and denied custody more than Parent A. Thus if P_c and P_r represent the percentage of participants who chose and rejected an option, respectively, then for Parent A $P_c + P_r < 100$ and for Parent B $P_c + P_r > 100$ (Shafir, Simonson & Tversky, 1993). Shafir then stated that for Parent B, $P_c + P_r = 119$ “is significantly greater than the 100 we would expect *if* choosing and rejecting were complementary $z = 2.48, p < .02$ ” (Shafir, 1993, p. 549, italics added).

The problem with this is that in a between-subjects design there is no reason to expect choosing and rejecting to be complementary; and, comparing 119 to 100 only makes sense if awarding and denying should in fact be complementary. This assumption only makes sense in a within-subjects design. In a within-subjects design, where the same participants who choose who to award custody to also choose who to deny custody to, participants should realize that awarding and denying custody to the same person would be irrational. This assumption cannot be made between subjects, where participants are not shown both conditions.⁵

Within- and between-subjects hypotheses

We have now seen multiple examples of within- and between-subjects designs not testing the same hypotheses. If within- and between-subjects designs are frequently not interchangeable

⁵ It could be added that in addition to this, Shafir’s analysis is not an appropriate use of a z test nor is calculating $P_c + P_r$ an appropriate way to analyze the data. As they do not pertain to the immediate discussion, these points will be discussed later in Experiment 1.

because they commonly test different hypotheses, it follows that there will be instances in which certain hypotheses can only be appropriately tested by one or the other design.

Therefore, let us call a “within-subjects hypothesis” any hypothesis worded in such a way that it can only be tested when the same participants (or cases) are exposed to all relevant conditions, to all relevant independent-variable levels or combinations of levels of multiple independent variables. Thus, a within-subjects hypothesis regarding participant responses to stimuli is any hypothesis wherein the level and specificity of the information required to put the hypothesis, as framed, to adequate test, necessitates that the same participants experience all relevant stimulus conditions. This could be because of the vagaries of the hypothetical construct involved or because of certain assumptions made that necessitate within-subjects manipulation (such as assuming responses should be complementary).

For example, in research on the conjunction fallacy, showing between subjects that group *A* rated “Bill is an accountant who plays jazz for a hobby” as more likely than group *B* rated “Bill plays jazz for a hobby,” tests whether probability judgments between groups seem to conform to the conjunction rule; it does not in any way constitute a test of whether the representativeness heuristic leads people to violate the conjunction rule of probability when making the relevant comparison in question (Tversky & Kahneman, 1983). The latter hypothesis can only be tested within subjects.

The disjunction effect and the conjunction fallacy are both examples of within-subjects hypotheses. If a hypothesis posits the existence of a “bias” that causes a person to answer in a specified manner when making a specific stimulus comparison, then the only design that can be used to test this hypothesis is one that in fact allows participants to make the actual comparison specified by the hypothesis. This requires that a within-subjects design be employed. If by what

we mean by a “conjunction effect” is merely that representativeness causes group responses to differ from each other, then the between-subjects test would be appropriate. This, in other words, would be a “between-subjects hypothesis.” This is not, however, what is meant by the “conjunction fallacy.” What is usually claimed (or implied) is that the representativeness heuristic causes the same people to violate the conjunction rule of probability. This is an entirely different hypothesis. It is also a stronger claim.

Further, these two claims, the stronger and weaker claims regarding representativeness, cannot imply the same conclusions. Tversky and Kahneman (1983) argued that because of this, when tested between subjects, violations of the conjunction rule can be called an “error,” but not a “fallacy.” It is only when tested directly within subjects that violating the conjunction rule of probability can be called the conjunction fallacy (Tversky & Kahneman, 1983). Tversky and Kahneman (1983) said that the reason for this is the increased transparency of the within-subjects presentation of stimuli. The problem with this assumption is that the conjunction fallacy is a within-subjects hypothesis regardless of whether or not a within-subjects design of it is in fact transparent. It is a within-subjects hypothesis because of the manner in which it is operationally defined. Though Tversky and Kahneman did not directly test for the transparency of their within-subjects designs, an important argument that they elaborately posed is that within-subjects designs may vary in their degree of transparency.

Are within-subjects designs a continuum of transparency?

If within-subjects designs vary in their degree of transparency, as Tversky and Kahneman (1983) suggested, this implies that it is not entirely accurate to think of between- and within-subjects designs as a dichotomy. Thus, the difference between them is more like a continuum: There is the between-subjects designs at one end and then a whole continuum of within-subjects

designs. Opposite the between-subjects design would be a “pure” within-subjects design. A study can fall at either end of this continuum (being a between- or a pure within-subjects design) or can fall somewhere in between, being neither a between-subjects design nor a pure within (but still technically being within subjects). In a pure within-subjects design, all relevant stimuli are displayed to the same participants at the same time and no attempt is made to in anyway lessen the design’s transparency. We start to move away from the pure within-subjects design whenever some action is taken to reduce the design’s transparency.

Take for instance a within-subjects questionnaire that contains questions relevant to three different experiments. If the questions pertaining to the three different experiments are not grouped by experiment and their order is mixed, which is a common practice in JDM research, then it is possible that this design is now less transparent than a pure within-subjects design and more transparent than a between. Perhaps we move even further away from the pure within-subjects design when there are time delays used between conditions, as in Tversky and Shafir (1992). Or what if one conducts a within-subjects study with three conditions, but has participants wait a month between each condition? This raises an interesting question: If participants wait so long between conditions that they do not accurately remember the previous ones when being tested again, then what is the real difference between this and a between-subjects design? If a pure within-subjects design and a between-subjects design represent the polar opposites of this continuum, how great is the difference of transparency between them regarding standard JDM experiments? All of this, of course, assumes degrees of transparency on the part of the within-subjects design. Experiments 2 and 3 below will also include a test for sensitivity along this proposed within-subjects continuum of transparency.

It might seem contradictory to doubt the transparency of within-subjects designs and then speculate about their degrees of transparency. If so, I would remind the reader that, as stated above, it is not being argued that within-subjects designs are not more transparent than between. What is being argued is that their actual transparency in the field of JDM is often exaggerated and is typically not of a magnitude that would impact the results of a study.

Between- vs. within-subjects designs: Other issues and controversies

There are many other differences between within- and between-subjects designs other than what we have so far discussed. There are, for example, reasons for preferring within-subjects designs other than the between-subjects context effects discussed above. Within-subjects designs require far fewer participants, and so their wider adoption would speed up the research process as well as reduce costs. In certain industrial or corporate settings where the information imparted in an experiment might be restricted or confidential, within-subjects designs would obviously be preferred as they minimize the number of people exposed to such information. Within-subjects designs are also substantially more powerful.

While these are certainly valid points, there are also downsides to the within-subjects design. As discussed above, the main counterpoint that some would make is likely that within-subjects designs are transparent and therefore contain demand characteristics. This assumption will be directly tested below in regards to JDM research. Other concerns involve the possible presence of within-subjects context effects, such as practice, fatigue, expectancy, order and carry-over effects. Thompson and Campbell (2004), for instance, pointed out that most research on deductive reasoning is done within subjects and investigated whether between-subjects designs might be more appropriate due to the possibility of expectancy and fatigue effects in within-subjects studies.

Thompson and Campbell tested for expectancy effects by varying between subjects how many problems participants solved. For example, participants might try harder when solving one reasoning problem between subjects than when solving eight within. They found that expectancy effects were negligible. Performance did change slightly as a function of the number of problems solved. This change, however, was very small and nonsignificant, thereby providing evidence that in research wherein participants solve eight reasoning problems back-to-back (eight being the median number of problems used in research on deductive reasoning), fatigue effects will likely not jeopardize the interpretation of within-subjects designs. Thompson and Campbell (2004) concluded that within-subjects designs should be preferred due to the substantial increase in power they afford. Zimmerman (1997), for instance, has pointed out that even when a correlation is as small as .05, using a between- rather than within-subjects design significantly reduces power.

Greenwald (1976) argued that though within-subjects designs are often preferred because of their ability to remove participant variance from error terms, this statistical efficiency is outweighed by the danger of practice and carry-over effects. Though he acknowledged the existence of between-subjects context effects, he seems to have greatly underestimated their prevalence and impact (Birnbaum, 1999; Parducci, 1965). Though with some treatments (such as medication), carry-over effects would doubtlessly be a concern, it is unlikely that they are an issue with the standard treatments used in JDM research. Similarly, most questionnaires utilized by JDM researchers are likely of a type and length that practice and fatigue effects would also not be a concern.

A concern that might be applicable in the field of JDM, however, is the possibility of order effects. In this paper, Experiment 2 will include a direct test for the influence of within-

subjects order effects. It should be added that, unlike between-subjects context effects, order effects are very easy to deal with through counterbalancing. Also, if one counterbalances the order the conditions of an experiment are presented in, the act of doing so renders most between-subjects designs unnecessary anyway. It does so by generating a dataset that can be analyzed both within and between subjects. Thus, if both analyses are appropriate this immediately allows one to treat design type as an additional independent variable. If one group sees two conditions in order A, B and the other in order B, A, analyzing all responses provides within-subjects data. Analyzing only the first condition presented for each group provides between-subjects data (though it also cuts the sample size in half). (It should be added though that if for a particular between-subjects design stimulus and context are confounded, that if one counterbalances conditions within-subjects and then analyzes the data between subjects as just described, the between-subjects context effect will still be present.) Experiment 2 will contain a survey counterbalancing the order of conditions (testing for order effects as well as comparing within- vs. between-subjects data) and will also contain a survey wherein the order of questions within a condition are counterbalanced.⁶

A very common (and sometimes strong) objection to the use of within-subjects designs is the argument that between-subjects designs more closely resemble the real world and therefore have greater ecological validity (Fischhoff, Slovic & Lichtenstein, 1979; Kahneman & Frederick, 2005). One must be specific in making this argument though, as others have argued that for many phenomena, within-subjects designs actually possess greater ecological validity than

⁶ If there is more than one question that participants must answer in each condition, then there is a difference between counterbalancing the order of questions within each condition and counterbalancing the order of the conditions themselves. In order to generate data that can be analyzed between subjects the order of the conditions themselves must be counterbalanced.

between (e.g., Greenwald, 1976). Pertaining to the field of JDM, the argument is that in the real world, people are generally not exposed to all of the relevant stimulus conditions of a judgment or decision and so a between-subjects design more closely mimics the conditions of real life.⁷ If the generalizability of a study's findings is based on how similar the experimental environment is to the natural environment that is being studied (even this point is controversial, see Birnbaum, 1999), and if between-subjects designs more closely resemble such natural environments, then presumably between-subjects designs are more generalizable. This point was made quite succinctly by Kahneman and Frederick (2005):

Between-subjects and factorial designs often yield different results in studies of intuitive judgment. Why should we believe one design rather than the other? The main argument against the factorial design is its poor ecological validity.... The between-subjects design, in contrast, mimics the haphazard encounters in which most judgments are made and is more likely to evoke the casually intuitive mode of judgment that governs much of mental life in routine situations. (pp. 280-281)

This concern is most easily answered. First, if the research hypothesis in question is a between-subjects hypothesis, and if between-subjects designs do in fact have greater ecological validity (since the real world typically presents only one set of values), then obviously the between-subjects design should be employed. If, however, the research hypothesis in question is a within-subjects hypothesis, then it is irrelevant if between-subjects designs more closely resemble the real world. A wholesale sacrificing of construct validity for a lesser and highly

⁷ An interesting point is that though life often only presents one set of values, within-subjects data (though perhaps of a limited scope) is often sought by people when making real-life decisions, and that further, when knowledge of alternative outcomes is unavailable, people typically attempt to generate a proxy of within-subjects data through counterfactual reasoning or mental simulation (Dunning & Parpal, 1989; Kahneman & Tversky, 1982).

debatable concern over ecological validity is to entirely fail to test the hypothesis at hand. If one fails to test the hypothesis at hand, then ecological validity is a moot point anyway.

Second, in the field of JDM where the dependent variable is typically a subjective judgment, between-subjects designs may not even provide a valid comparison. If context and stimulus are confounded, the resulting data will be misleading and useless. Thus, Kahneman and Frederick's (2005) above argument ignores Birnbaum's (1999) and Parducci's (1966) point that all judgments are relative to the context in which they are made, and if the contexts in which they were made are different between groups, the judgments cannot be meaningfully compared. It does not make sense, for example, to try and save a between-subjects finding by arguing that, such as is the case with the base-rate neglect fallacy, between-subjects designs are to be preferred since their findings are more "generalizable." It does not matter if between-subjects studies of the base-rate neglect fallacy have higher ecological validity, if they do not even allow one to measure the phenomenon in question. Birnbaum (1999), for instance, argued that this is in fact the case, adding that an assessment of the within-subjects studies of the base-rate neglect fallacy suggests that this "fallacy" does not actually exist.⁸

Introduction conclusions

As stated above, though Birnbaum's (1982, 1999) argument regarding between-subjects context effects is vitally important and likely detrimentally underappreciated, the author believes that he may be overstating his case somewhat. It might not be necessary to sweepingly conclude that if the dependent variable is a subjective judgment that between-subjects designs should

⁸ These arguments are highly related to Brunswik's work. Birnbaum (1999) pointed out that many use Brunswik's (1956) work to argue against the use of within-subjects designs, while ironically ignoring that Brunswik actually argued against the use of between-subjects designs!

never be used. Assume, for a moment, that Birbaum is completely correct. This, combined with the author's above argument regarding between- and within-subjects hypotheses, would leave a great many studies in an awkward position, as it would result in the following matrix:

		Dependent Variable	
		Objective	Subjective
Hypothesis	Within	Use a within-subjects design. Examples: certain longitudinal medical studies, giving participants a standardized test before and after training of some kind, etc.	Use a within-subjects design. Examples: testing for the conjunction fallacy, testing for disjunction effects, etc.
	Between	Use a between-subjects design. Example: medical studies utilizing a treatment and placebo group that have an objective measure such as temperature (fever), etc.	Neither the standard within- nor between-subjects design can validly be used. One possibility is a triple-blind between-subjects study.
Looking at the matrix			
above, it would seem that a			great

many JDM studies fall into the lower-right cell. This number could be lessened, however, by considering whether the hypothesis to be tested is really a between- or within-subjects hypothesis. If a between-subjects design was used when a within-subjects design would have been more appropriate (such as is the case with the conjunction fallacy, the disjunction effect and possibly base-rate neglect), then the study should rather fall into the upper-right cell. The author, however, does not wholly agree with the total abandonment of the between-subjects design for the investigation of heuristics and judgment. The aim, rather, should be on raising the awareness of between-subjects context effects, as they are likely more common and a more severe confound than is appreciated by researchers, both basic and applied.

Deciding upon design type then is itself a critical thinking exercise. It is a decision that requires careful thought which cannot be automated by running a simple statistic or looking at a

table. One must look carefully at one's hypotheses. How are the hypothetical constructs operationally defined? Do such definitions lend themselves better to between- or within-subjects designs? (Or perhaps both work.) If between, are there any conceivable context effects that could be present? (One must be careful here, however, as one of Birnbaum's most important arguments is that between-subjects context effects are likely present even when unrecognized.) Such questions are seldom answered—let alone thought about—when selecting design type (Greenwald, 1976).

In sum, it must not be forgotten that sometimes the results of a study are simply a function of the experimental protocol employed (Greenberg & Lambdin, 2007); that between- and within-subjects designs often yield different results for reasons other than task transparency (Birnbaum, 1982); and further, that sometimes the operational definition of the construct to be explored precludes any meaningful use of between-subjects designs (Lambdin & Burdsal, 2007). Such points will hopefully help make salient the bidirectional relationship between what is being studied and the research design that is adopted to study it. The choice of research design used is based on an understanding of the nature of that which we wish to study. What we find, however, is also a function of the methodology adopted, the operational definitions used, the design type decided upon, the level at which the data are aggregated as well as the particular analyses executed. The nature of any scientific “finding,” in JDM, psychology or otherwise, is not properly understood without an appreciation of this codependent relationship. This applies also to the hard sciences, but may apply doubly so to all fields within the social sciences. Social science data is sometimes said to always be inherently subjective in nature (Loftus, 1991).

It was stated that there are two assumptions that can be found in JDM literature concerning between- and within-subjects designs. The first was the *a priori* assumption that

presenting stimuli within subjects renders the experimental task transparent to the point of confounding the results. The second was that—transparency aside—between- and within-subjects designs are interchangeable. Much of the preceding discussion was spent addressing the second assumption. The first assumption will now be put to empirical test throughout the course of the four experiments that follow. The purpose of Experiment 1 is to see if two famous between-subjects experiments will replicate within subjects and to see if the qualitative data collected reveal any special insight into the task as a result of showing all conditions to the same participants.

CHAPTER 2

EXPERIMENT 1

Introduction

Experiment 1 is the first of four experiments designed to address the first assumption discussed above: that presenting stimuli within subjects renders the experimental task transparent to the point of threatening the integrity of the results. In Experiment 1, two between-subjects experiments from the JDM literature will be replicated within subjects. The two experiments replicated are Shafir's (1993) child custody case experiment and Tversky and Kahneman's (1981) famous Asian disease problem experiment. These two experiments were chosen because it was thought that many would assume that both of them should be highly transparent when presented within subjects.

In Shafir (1993), participants 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 all around more average attributes, whereas Parent *B* (Shafir assumed) had both some really good and some really bad attributes. Shafir hypothesized that, because of the theory of reason-based choice, when asked which parent to award custody to participants would selectively search for reasons to award custody. Since Parent *B* has more extreme positive attributes than *A*, most participants would award custody to Parent *B*. Similarly, when asked which parent they would deny custody to, participants would selectively search for reasons to deny custody. Since Parent *B* has more extreme negative attributes than *A*, most participants would deny custody to *B*. The hypothesis, therefore, is that most participants will award and deny custody to Parent *B*.

The second experiment replicated here is Tversky and Kahneman's (1981) Asian disease problem. Tversky and Kahneman presented one group two programs and another

group two different programs. The “programs” were proposed solutions to a hypothetical outbreak of an “Asian disease.” We will call the “survival” frame Programs *A* and *B* (worded in terms of how many people would survive if the program in question was adopted) and the “mortality” frame Programs *C* and *D* (worded in terms of how many people would die if the program in question was adopted). Tversky and Kahneman (1981) hypothesized that a preference reversal would occur between the two frames, a framing effect. Framing effects occur when different descriptions of functionally equivalent information cause people’s preferences to differ.

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. The same information, presented in two different fashions, can cause two different subjective reactions. This is in part due to the value function of prospect theory (Kahneman & Tversky, 1979), which states that people are more likely to accept a smaller, sure gain than gamble for greater gain, whereas they will more likely attempt to gamble out of greater loss rather than accept a smaller, sure loss. This is also in part due to fact that different frames often make different attributes salient. People will gladly accept a discount for paying a bill early, but do not like a late fee for paying late, even though this amounts to the same outcome. Similarly, a person might favor ground beef that is 75% lean over beef that is 25% fat.

In addition to replicating these between-subjects experiments using a within-subjects design, participants were asked to describe the thoughts and/or impulses they experienced while answering the questions for each replication in an open-ended response. If a participant does not

respond to the stimuli in the hypothesized fashion, it was thought that the data obtained from these open-ended responses might reveal that they initially felt an impulse to respond as hypothesized but then recognized within subjects that doing so would be irrational. This would constitute evidence of task transparency. Furthermore, if a great many participants comment on why answering or not answering the way they did struck them as odd, this would indicate an insight into the research hypothesis for that experiment. This is the case since what is being hypothesized for each experiment is that participants will answer in a normatively irrational manner.

Method

Participants

Seventy-eight undergraduate students at Wichita State University volunteered to participate for extra credit. One participant's answers were deleted since the typed responses were random letters, resulting in a sample size of 77.

Procedure

Experiment 1 replicates two experiments. Data from these studies were originally obtained via between-subjects designs. Each of the experiments had two versions of a scenario which were presented to two different groups of people. In this experiment, respondents were asked to make judgments about both versions of the scenario, for both experiments. It has been argued that mixing up the questions in a survey that pertain to different experiments reduces task transparency, presumably because participants will have a harder time identifying what the relevant comparisons for each experiment are (e.g., Tversky & Shafir, 1992). The problems pertaining to each experiment in this replication were not randomized. This was done, assuming the above argument is correct, in order to maximize task transparency.

After each replication, participants were asked to write whether anything struck them as odd about their responses. If presenting conditions within subjects renders the experimental task transparent, then participants in within-subjects replications should possess insight into the research hypothesis being tested. For each experiment, what is being hypothesized is that participants will answer in normatively irrational ways. Therefore, if presenting stimuli within subjects renders the design transparent, then participants should become aware, within subjects, that they are being expected to respond to the stimuli illogically.

Materials

The experiments replicated were Shafir's (1993) custody trial experiment and Tversky and Kahneman's (1981) famous "Asian disease problem." Participants were first presented with Shafir's (1993) custody trial experiment. The following instructions were provided:

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?

Participants were then presented with Parents *A* and *B*:

Parent A

Average income

Average health

Average working hours

Reasonable rapport with the child

Relatively stable social life

Parent B

Above-average income

Very close relationship with the child

Extremely active social life

Lots of work-related travel

Minor health problems

Participants were asked to identify the parent to whom custody should be awarded. After making their choice, they were presented with the following additional instructions:

We would like you to now answer an alternative version of the question you just completed. It is important that you pretend that you did not just read the question you just answered. Act as though this is the first time you've answered this question, answering as honestly as possible.

Please answer with whatever your first impulse is.

Participants were then presented with the same scenario replacing the word “award” with “deny.” Participants were asked to identify the parent to whom custody should be denied. After they provided a second judgment about custody, the participants were asked to describe their thought process in an open-ended response.

Participants were then presented the “Asian disease problem” (Kahneman & Tversky, 1981). The following instructions were provided:

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.

Participants were asked to indicate which program they would like to adopt. After making their choice, they were given the following instructions:

We would like you to now answer an alternative version of the question you just completed. It is important that you pretend that you did not just read the question you just answered. Act as though this is the first time you've answered this question, answering as honestly as possible.

Please answer with whatever your first impulse is.

Participants were then presented with the same instructions as above, but using two new programs:

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.

Participants were asked to indicate which of these two programs they would adopt. They were then again instructed to describe their thought process. All data were collected via computer using the software program, MediaLab.

Results

The child custody case

Shafir (1993) found between subjects that 36% of participants chose to award custody to Parent A, 45% denied custody to Parent A, 64% awarded custody to Parent B and 55% denied custody to Parent B ($N = 170$). Shafir then showed that for Parent B $P_c + P_r = 119$ and for Parent A $P_c + P_r = 81$. Shafir argued that for Parent B, $P_c + P_r = 119$ is significantly greater than the 100 that we would expect if choosing and rejecting were complementary, $z = 2.48, p < .05$ (Shafir, 1993; Shafir, Simonson & Tversky, 1993). Again, comparing 119 to 100 only makes sense if we really can expect that choosing and rejecting should be complementary, and, in a between-subjects design there is no reason whatsoever to expect that they should be. This assumption only makes sense in a within-subjects study. Within subjects, what is one person awarding and denying custody to the same parent is between subjects two different people disagreeing over who should have custody of the child.

Yet another problem with Shafir's conclusion is that in addition to inappropriately aggregating his data, he did not properly use the z test that he is relying on to claim a significant result. A z test cannot be used to compare 119% to 100%. Since the data are binary there is no mean or standard deviation and so a z test for a mean cannot be used. A z test for two proportions also cannot be used because 119 is not a frequency; it is the sum of two percentages

and it exceeds 100%. Furthermore, since $P_c + P_r$ can be > 100 for Parent B when B is not awarded and denied custody more than A , showing that $P_c + P_r > 100$ for Parent B cannot be a meaningful demonstration. Again, Shafir (1993) found that for Parent B $P_c + P_r = 119$. If 37% of participants award custody to Parent B and 80% deny custody to B , $P_c + P_r = 119$ even though participants did not award and deny custody to Parent B more than A . Because of this, the results of the replication will be analyzed differently than Shafir (1993). Below are the results of the present within-subjects replication.

Within-subjects replication data, the child custody case (N = 77)

Award A	Award B	Deny A	Deny B
52% (40)	48% (37)	52% (40)	48% (37)

Figure 1. Within-subjects replication data, the child custody case

Twenty-five percent of the sample (19 participants) awarded and denied custody to the same parent, though, as seen above, more participants awarded and denied custody to the impoverished option (Parent A) than to the enriched option (Parent B). This is not in the direction of Shafir's prediction; therefore, this data does not require further analysis.

Nevertheless, a quarter of all participants (25%) awarded and denied custody to the same parent. The 95% confidence interval for that 25% (.25) was .15 to .34. Since zero does not fall within this range, the number of participants who answered illogically (19) was significantly greater than zero.

Since what was being hypothesized was that the majority of participants would irrationally award and deny custody to Parent *B*, the open-ended comments were examined for evidence that participants recognized this or reported feeling an impulse to do so or exhibited an understanding of the manipulation. Only 9% of comments (seven comments out of 77) betrayed an understanding that the responses made should be complementary, that one should not award and deny custody to the same parent.

The Asian disease problem

Tversky and Kahneman (1981) found between subjects that 72% of participants chose Program *A*, 28% chose *B* (for *A* and *B*, $N = 152$), 22% chose *C* and 78% chose Program *D* (for *C* and *D*, $N = 155$). Below are the results of the present within-subjects replication of Tversky and Kahneman (1981).

Within-subjects replication data, the Asian disease problem (N = 77)

Program <i>A</i>	Program <i>B</i>	Program <i>C</i>	Program <i>D</i>
------------------	------------------	------------------	------------------

77% (59)	23% (18)	40% (31)	60% (46)
----------	----------	----------	----------

Figure 2. Within-subjects replication data, the Asian disease problem

When running this study between subjects, no one actually switches preferences between frames since no one is exposed to both conditions. Running the study within subjects, then, allows one to look at this interesting statistic. Here, 36% of participants switched preferences between frames. Of this 36%, 100% chose Programs A and D. The 95% confidence interval for this 36% is .26 to .47, indicating that it was significantly greater than zero. The within-subjects replication was successful, though Tversky and Kahneman's original between-subjects study found a more dramatic preference reversal. To illustrate, they found that 78% of participants chose Program D, whereas here 60% chose D.

With a sample size of 77, there is 80% power to detect 20 percentage-point differences using two-tailed z tests for two proportions. In the present within-subjects results, 77% of participants preferred Program A to B. This 77% was significantly greater than 50% (what would be expected if there was no preference between programs), $z = 3.48$, $p < .05$. In the other condition the majority of participants, 60%, preferred Program D. Though this 60% was not significantly different from 50%, $z = 1.25$, $p > .05$, it is nevertheless true that the same qualitative

pattern of data was found within subjects as between. The open-ended comments were also examined for evidence that participants recognized that they were being expected to answer irrationally as well as for evidence that they had caught on to the manipulation. Here, 18 participants (23%) provided comments that showed such an understanding.

Discussion

The within-subjects replication of Shafir (1993) was not successful. The results did not support the hypothesis offered by the theory of reason-based choice, namely that the enriched option would be both awarded and denied custody more than the impoverished option. However, the replication of Tversky and Kahneman (1981) was successful. Presenting all programs to the same participants did not greatly alter the results. In addition, few of the open-ended comments betrayed an understanding of the research hypothesis or experimental manipulation. Instead, almost all comments were rationalizations of the choices made, such as, “I like to go for the underdog.” In the child custody case, only seven comments (9%) betrayed an understanding that the decisions should be complementary. For example, one participant wrote: “I think it’s odd that I choose the same parent to have the child and again to not have the child.” The other 81% of comments consisted mostly of rationalizations for the choices made, such as stating that they wanted what was best for the child. In the Asian disease problem, 23% of participants showed some understanding of the research hypothesis or experimental manipulation. Some examples include: “It did strike me as slightly odd that i would choose the exact opposite program in each of the questions simply based on the wording used” (sic) and “Basically the option decribe the same things” (sic). As with the child custody case replication, most comments were rationalizations of the choices made, such as stating that they wanted to

save as many lives as possible. Few people wrote more than one or two sentences, most writing no more than seven or eight words.

Interestingly, for both of the replications, some of the comments suggest that participants interpreted the stimuli differently than how the experimenter intended them to be interpreted. If participants' interpretations of the stimulus material by-and-large do not gel with how the experimenter assumes they will interpret it, construct validity may be lacking. In the child custody experiment, participants did not seem to agree with the experimenter's assumption that Parent *B* had both more negative and positive attributes. For many participants, Parent *B*'s negative attributes outweighed the good ones. Nineteen percent of participants explicitly stated Parent *B* is an unfit parent simply because s/he has an "extremely active social life." For example: "I was thinking that in Parent *A*, the social life is minimal...Parent *B* on the other hand...have a very active social life" (sic) and "Parent *A* had a stable social life."

In the Asian disease problem, Tversky and Kahneman (1981; 1986) assumed that Programs *A* and *C* and then *B* and *D* describe numerically equivalent results. Given this assumption, the fact that in the survival frame the majority of participants preferred Program *A* whereas in the mortality frame the majority preferred *D* (rather than *C*, which is assumed to be numerically equivalent to *A*) would constitute a "framing effect." As with the child custody case experiment, however, some of the comments suggest this experiment might lack construct validity and call into question the manipulation. This supports Kuhberger's (1995) argument that the outcomes in the Asian disease problem are not adequately specified and are therefore open to more than one interpretation.

It only makes sense to call switching from Program *A* to *D* or from *B* to *C* a "framing effect" if it can be assumed that doing so is to answer inconsistently. In other words, the

experimenter must assume that the participants assume that Programs A and C and Programs B and D are numerically equivalent statements, that saying “400 people will certainly die” implies that “200 people will certainly live.” Nine percent of the participants, however, indicated in their comments that they did not see these statements as equivalent outcomes, which is actually a valid interpretation. Such comments included: “Program A can save *at least* 200 people” (italics added) and “C can guarantee 400 deaths but not 200 survival” (sic). Thus, given this interpretation, saying “400 people will certainly die” does not imply that “200 will certainly live,” but rather that “one can be certain at least 400 out of 600 will die, maybe 450, maybe 575 will die.”

Conversely, “200 will be saved” was taken by some to mean “with Program A at least 200 will live,” which does not tell one how many will in fact die. This means in Program A at least 200 will be saved (maybe 500 will live) and with *D* one does not know that anyone at all will certainly die. Therefore, given this interpretation, it is not only rational to choose *A* and *D* but it also does not constitute a framing effect in any way, shape or form. Twenty-eight participants (36% of the sample) answered “*A*” and “*D*.” This is significantly different from zero as the 95% confidence interval for this 36% is .26 to .47. Interestingly, 100% of participants who “switched” answers chose programs *A* and *D*. Nobody chose *B* and *C*.

A potential limitation of this study is the instructions presented to the participants. Recall that between conditions participants read the following: “It is important that you pretend that you did not just read the question you just answered. Act as though this is the first time you’ve answered this question, answering as honestly as possible. Please answer with whatever your first impulse is.” First of all, such instruction could possibly be confusing. How can participants act as though they had not read the first question when research on the hindsight bias

demonstrates our inability to disregard prior knowledge? (e.g., Fischhoff, 1975) This is similar to asking jury members to not consider something that has already been said. Secondly, these instructions create the threat of a strong demand characteristic to answer irrationally even if they recognized the problem with doing so. For these reasons, these experiments were replicated again in Experiment 2, sans any instruction regarding the participants' responses.

Another possible limitation of this study regards the open-ended question used. It was designed to look for an understanding of the research hypothesis in question but did not explicitly enough ask for such understanding. Therefore, the qualitative data obtained may actually underestimate the percentage of participants who obtained insight into the task within subjects. This possibility will be addressed in Experiment 3, in which participants will be explicitly asked to try and figure out the research hypothesis for each experiment replicated within subjects.

CHAPTER 3

EXPERIMENT 2

Introduction

The purpose of Experiment 2 is to expand upon the results of Experiment 1, replicating a greater number of studies within subjects without the use of any leading instructions. In Experiment 2 there are two surveys, each counterbalanced differently. In the first survey, four experiments are replicated within subjects. The order of the options participants chose from in each condition was counterbalanced to test for order effects. It had been suggested by a reviewer that in Experiment 1 most people chose the impoverished option simply because s/he was always presented as Parent A.⁹ Thus, for the first survey, the attributes of Parent A and Parent B were switched for half of the questionnaires. For the Asian disease problem, this means that for half the participants the programs were presented normally, and for the other half, the descriptions of Programs A and B were switched, as were the descriptions of Programs C and D.

This survey was also designed to test whether answers vary as one moves along the within- to between-subjects design continuum discussed above. Recall that it is likely not entirely accurate to think of the difference between a within- and between-subjects design as a dichotomy. This is because the within-subjects design represents a sort of continuum. A “pure” within-subjects design can be thought of as one in which no attempt is made to lessen task transparency. Thus, in a pure within-subjects design, questions or stimuli relevant to different experiments conducted in the same survey would be grouped together and presented at the same time. We move away from this when questions pertaining to different experiments in the same

⁹ Though, if this explanation were correct it would not make sense that in Experiment 1’s Asian disease problem replication the majority of participants chose Program D rather than C.

survey are mixed or randomized (as done in Bastardi & Shafir, 1998), or when time delays are introduced in between within-subjects conditions. Tversky and Shafir (1992), for example, in one experiment had participants wait at least a week in between the conditions of a within-subjects experiment. To test for differences along this continuum, for half of both the “order normal” and “order reversed” questionnaires, the conditions for each scenario were presented on the same page. For the other half, they were presented on different pages. This resulted in four versions of the first survey. Thus, in the Asian disease problem, half of participants saw Options *A* and *B* and Options *C* and *D* on the same page. For the other half, Options *A* and *B* were shown on one page and then Options *C* and *D* were shown on another.

The second survey replicates three of the same experiments as the first and was counterbalanced so that the data could be analyzed both within and between subjects. Replicating a between-subjects study within subjects is one way to test whether the data changes when switching from one design to another. A far more direct way is to set up a within-subjects design so that its results can be analyzed both within and between subjects. This allows for the same issue to be addressed using the same participants, thereby removing the influence of using different samples on the data.

In order to set up the survey so that it can be analyzed both within and between subjects, the order of conditions themselves were counterbalanced. (This also provides an additional test for order effects). For example, in the Asian disease problems, half of the participants saw the survival frame first and the other half saw the mortality frame first.

Two of the experiments replicated in Experiment 2, Tversky and Kahneman’s (1981) Asian disease problem experiment and Shafir’s (1993) child custody case experiment were discussed in Experiment 1. The other two experiments replicated here are Tversky and

Kahneman's (1986) marbles lotteries experiment and Bastardi and Shafir's (1998) Princeton applicant experiment. As mentioned in Experiment 1, the child custody case and Asian disease problem experiments were chosen because it was thought that many researchers would assume that both should be highly transparent when presented within subjects. Here this also applies to the Princeton applicant experiment. The marbles lotteries was included as an experiment that would seem to be less transparent within subjects. Here we shall see how these assumptions bear out on the actual replication results. (The results from Experiment 1 were likely somewhat confounded by the wording of the instructions.)

In Tversky and Kahneman's (1986) marbles lotteries experiment, one group of participants were shown Options *A* and *B* below, and another group of participants were shown what we will call Options *C* and *D*. In each group, participants were asked which of the two lotteries they would prefer to play:

___ 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

In Tversky and Kahneman's (1986) study, participants were told that 1/10 of them would actually play the lotteries chosen. In their experiment, for the second group of participants the two lotteries were still entitled "Option *A*" and "Option *B*." In the present within-subjects replication, the two lotteries in the second condition were renamed "*C*" and "*D*." For this discussion we will keep the *C* and *D* to help distinguish between the four lotteries.

Between Options *A* and *B*, *B* is the dominating lottery. The dominating lottery is the one with the better odds of winning out of the two. 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 participants would choose Option *C* over *D*, even though Option *D* is the same as *B* and dominates *C*.

In the other new study to be replicated, Bastardi and Shafir (1998) presented participants with a description of a hypothetical applicant to Princeton. Participants were asked to imagine they were on the admissions committee. For one group of participants, the applicant had a high school average of B. These participants were asked whether they would accept or reject the applicant. For the other group of participants, there were conflicting reports regarding the high school average, with one report stating that it was A and another stating B. These participants were asked whether they would accept the applicant, reject the applicant or choose to wait until the grade discrepancy could be cleared up. Those who chose to wait were informed that the high school average was in fact B and were then asked whether they would accept or reject the

applicant. The type of counterbalancing employed in the second survey in Experiment 2 does not, however, work for this replication. What is of interest here, within subjects, is how many people who already accepted the applicant with a B average then elect to wait for what should then be obviously irrelevant information. The conditions in that experiment must, therefore, be presented in a particular order. Because of this, the replication of Bastardi and Shafir (1998) will only be discussed for the first survey.

Method

Participants

Two surveys were administered for Experiment 2. In the first, 96 undergraduate students at Wichita State University volunteered to participate for extra credit. One participant was dropped since s/he did not answer all of the questions, resulting in a sample size of 95. In the second survey, 89 undergraduate students at Wichita State University volunteered to participant for extra credit.

Procedure

Participants were undergraduate psychology students. They were asked, in class, if they would like to participate in an experiment for extra credit. Anyone who did not want to participate was free to opt out, and anyone who did not want to participate but still wanted the extra credit was provided other ways of obtaining it. Those who chose to participate were given a paper-and-pencil survey that took, on average, 10 minutes to complete.

Materials

The experiments replicated in Experiment 2 are: the Asian disease problem from Tversky and Kahneman (1981), the child custody case from Shafir (1993) (both discussed in Experiment

1), the marbles lotteries from Tversky and Kahneman (1986) and the Princeton applicant scenario from Bastardi and Shafir (1998) (both discussed below).

It was important that the instructions used for Experiment 2 not create any demand characteristics. Therefore, the coercive language used in Experiment 1 was removed. The new instructions read:

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.

The instructions for the Asian disease problem and the child custody case were the same as in Experiment 1, excepting that between conditions no additional instruction was given at all; i.e., participants were not in any way asked to “answer with their first impulse” or “ignore the last question,” etc. (A complete questionnaire for Experiment 2 with full instructions used can be seen in Appendix A.) The marbles lotteries read as follows:

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

And the Princeton applicant scenario read as follows:

Imagine you are on the admissions committee of Princeton University. You are reviewing the file of an applicant who plays varsity soccer, has supportive letters of recommendation, and is editor of the school newspaper. The applicant has a combined SAT score of 1250 and a high school average of B. Do you choose to:

____ Accept the applicant?

____ Reject the applicant?

Again, imagine that you are on the admissions committee of Princeton University. You are reviewing the file of an applicant who plays varsity soccer, has supportive letters of recommendation, and is editor of the school newspaper. The applicant has a combined SAT score of 1250 but you have two conflicting reports of the applicant's high school average grade. The guidance counselor's report indicates a B average, while the school office reported an A average. The school has notified you that the records are being checked, and that you will be informed within a few days which of the averages is the correct one. Do you choose to:

____ Accept the applicant?

____ Reject the applicant?

____ Wait for clarification from the applicant's school before deciding?

If you chose to wait for clarification in the question above, please answer the following:

The school informs you that the applicant's average grade is a B. Do you choose to:

____ Accept the applicant?

____ Reject the applicant?

All data were collected using paper-and-pencil questionnaires.

Results

First survey (N = 95)

Testing for order effects

This survey includes a test for the impact of order effects. The purpose of the replications is to see whether the overall patterns of data will be similar between designs. The point, therefore, of testing for order effects here 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. Results will of course fluctuate from replication to replication, but only a substantially large percentage-point difference would change the results of the study. The first survey has a sample size of 95, with 47 participants in one group and 48 in the other. This yields 80% power to detect differences of 20 percentage points using two-tailed z tests for two proportions, more than adequate power given that we are not here interested in percentage-point differences much smaller than this.

To test for order effects in the marbles lotteries problem (Tversky & Kahneman, 1986), half the participants were given Options *B* and *D* as the dominating (better) lotteries and the other half were given Options *A* and *C* as the dominating lotteries. Here are the results for Tversky and Kahneman's (1986) marbles lotteries experiment:

The marbles lotteries, first (n = 48) vs. second (n = 47) option presented as dominating lottery

Option A	Option B	Option C	Option D
----------	----------	----------	----------

B and *D* dominating: 11% (5) 89% (42) 79% (37) 21% (10)

A and *C* dominating: 92% (44) 8% (4) 27% (13) 71% (34)

Figure 3. The marbles lotteries, first vs. second option presented as dominating lottery

The results almost perfectly reverse themselves depending on whether the dominating lottery is the first or second of the two lotteries presented. Counterbalancing resulted in a three percentage-point difference with Options *A* and *B* and an eight percentage-point difference with Options *C* and *D*. Neither difference was significant, $z = .50, p > .05$ and $z = .88, p > .05$, respectively. For all data that follows regarding the marbles lotteries, the second options (*B* and *D*) will be treated as the dominating lotteries. (This is how they appeared in Tversky and Kahneman, 1986.)

The Princeton applicant, “accept”($n = 47$) vs. *“reject”* ($n = 48$) presented as first option

Presenting this scenario within subjects allows for a test of disjunction effects. What is of interest here, therefore, is how many participants irrationally choose to wait for irrelevant information after already accepting the applicant when the high school average was *B*. If you accepted the applicant when the high school average was *B*, then you should also accept the

applicant when the high school average is either B or A. Not doing so is a violation of Savage's (1954) sure thing principle (STP). A disjunction effect is simply a violation of STP (Tversky & Shafir, 1992).

Both when "accept" was presented as the first option and as the second, 83% chose to accept the applicant. When "accept" was presented as the first option, 30% of participants (14 participants) who accepted the applicant then chose to wait for irrelevant information. When "reject" was presented as the first option, 40% (19) chose to wait after already accepting the applicant. The 95% confidence interval for the proportion of participants who violated the sure thing principle in the "accept"-first group was .22 to .50. For the "reject"-first group it was .33 to .62. Thus, in both groups the proportion of participants whose choices violated the STP was statistically significant.

The child custody case, first (n = 48) vs. second (n = 47) parent presented as "enriched option"

	Award A	Award B	Deny A	Deny B
Parent A enriched:	40% (19)	60% (29)	63% (30)	38% (18)
Parent B enriched:	70% (33)	30% (14)	30% (14)	70% (33)

Figure 4. The child custody case, first vs. second parent presented as “enriched option”

The results almost perfectly reverse themselves depending on whether Parent *A* was presented as the “enriched” or “impoverished” option. (The enriched option has both more extreme positive and negative attributes. The impoverished option has more average attributes all around.) When the enriched option was presented first, 60% of participants awarded custody to the impoverished option vs. 70% when the impoverished option was presented first. This difference was not significant, $z = 1.00, p > .05$. Again, when the enriched option was first, 38% denied custody to the impoverished option vs. only 30% when the enriched option was second. This difference was also not significant, $z = .82, p > .05$. This suggests that the results of Experiment 1 were not due to an order effect but were rather the result of the leading instructions given to participants. For the remainder of the discussion, Parent *A* will be treated as the “impoverished” option and Parent *B* as the “enriched” option.

The Asian disease problem, first (n = 48) vs. second (n = 47) program presented as gamble

	Program A	Program B	Program C	Program D
Gamble is A & C:	33% (16)	67% (32)	65% (31)	35% (17)
Gamble is B & D:	60% (28)	40% (19)	30% (14)	70% (33)

Figure 5. The Asian disease problem, first vs. second program presented as gamble

Thirty-three percent of participants chose the gamble when it was Program A vs. 40% when it was B. This difference is not significant, $z = .72, p > .05$. Sixty-five percent chose the gamble when it was Program C vs. 70% when it was D. This difference was also not significant, $z = .52, p > .05$. For the remainder of the discussion, Programs A and C will be treated as the sure things and Programs B and D as the gambles.

As noted above, there is more than adequate power to detect moderate to large sized order effects. Small order effects, of course, may have been missed here. This is not a concern because the qualitative pattern of the data is the same for both presentations of stimuli in all four replications. Even though there may be some minor order effects present they are not of a magnitude to actually alter the pattern of the results. It is safe, therefore, to collapse across these conditions.

Testing for differences along the “within-subjects continuum”

It is argued in the Introduction that a within-subjects design is perhaps best thought of as a continuum, some being more transparent than others. Some researchers using a within-subjects

design take measures in attempt to lessen its perceived transparency, such as presenting questions on different pages or introducing time delays between the presentation of conditions. Here, results will be presented for the different- vs. same-page manipulation. The purpose is to see if simply presenting conditions on the same vs. than on different pages is enough to impact the results obtained. Though there are now 49 participants in one group and 46 in the other, there is still 80% power to detect differences of 20 percentage points. Below are the results for Tversky and Kahneman's (1986) marbles lotteries experiment.

The marbles lotteries, different (n = 49) vs. same (n = 46) page

	Option A	Option B	Option C	Option D
Same page:	11% (5)	89% (41)	74% (34)	26% (12)
Different pages:	8% (4)	92% (45)	76% (37)	14% (7)

Figure 6. The marbles lotteries, different vs. same page

The different- vs. same-page manipulation resulted in a three percentage-point difference for Options A and B. This difference was not significant, $z = .50, p > .05$. The two percentage-point difference for Option C was not significant, $z = .23, p > .05$. Fewer participants chose

dominating Option *D* (14%) than when shown on the same page (26%), resulting in a difference of 12 percentage points. This difference was also nonsignificant, $z = 1.47, p > .05$.

The Princeton applicant, different (n = 49) vs. same (n = 46) page

When conditions were presented on different pages, of the 80% (39 participants) who already accepted the applicant with an average of B, 46% (18) then chose to wait for noninstrumental information. When conditions appeared on the same page, of the 90% (40) who accepted the applicant with an average of B, 38% (15) then waited for irrelevant data. The 95% confidence interval for the 46% who violated the sure thing principle in the different pages group was .32 to .61. For the same page group it was .24 to .51. Thus, in both groups there was a statistically significant disjunction effect present.

The child custody case, different (n = 49) vs. same (n = 46) page

	Award A	Award B	Deny A	Deny B
Same page:	59% (27)	41% (19)	39% (18)	61% (28)
Different pages:	71% (35)	29% (14)	29% (14)	71% (35)

Figure 7. The child custody case, different vs. same page

For Award *A* and Award *B*, a 12 percentage point difference resulted from showing conditions on the same page vs. different pages. However, this difference was nonsignificant, $z = 1.23, p > .05$. The resulting 10 percentage-point difference for Deny *A* and Deny *B* was also not significant, $z = 1.03, p > .05$.

The Asian disease problem, different (n = 49) vs. same (n = 46) page

	Program A	Program B	Program C	Program D
Same page:	63% (29)	37% (17)	41% (19)	59% (27)
Different pages:	63% (31)	47% (18)	24% (12)	76% (37)

Figure 8. The Asian disease problem, different vs. same page

Presenting conditions on different vs. the same page did not create a difference for Program *A*. For Program *B*, the 10 percentage-point difference that resulted was not significant, $z = .99, p > .05$. For Programs *C* and *D*, a 17 percentage-point difference resulted and was also not significant, $z = 1.76, p > .05$.

There was, again, adequate power for detecting moderate to large changes in the data caused by this manipulation. It can be seen that presenting conditions on different pages did

slightly enhance the effects in question, though this impact did not in any way change the overall pattern of results and none of the differences created were statistically significant. It is therefore safe to also collapse across the different- vs. same-page conditions, allowing for all data from the first survey to be combined. Collapsing across the different- vs. same-page manipulation, as well as the order effect conditions, the replication results for the first survey are as follows:

First survey, replication data

Below are the replications results from the first survey. For the following analyses there is 86% power to detect differences of 20 percentage points.

The marbles lotteries

Tversky and Kahneman (1986) found that between Options A and B, 100% of participants chose dominating Option B ($N = 88$). In the other group however, 58% of participants chose dominated Option C ($N = 124$). Below are the replication results for the marbles lotteries experiment, all data from the first survey combined.

Within-subjects replication data, the marbles lotteries ($N = 95$)

Option A	Option B	Option C	Option D
9% (9)	91% (86)	76% (71)	24% (23)

Figure 9. Within-subjects replication data, the marbles lotteries

Looking at the present replication results, the 91% who preferred dominating Option *B* was significantly different from the 50% that would be expected if there was no preference between options, $z = 6.20, p < .05$. In the other condition, 76% then preferred the dominated Option *C*. This was also significantly different from 50%, $z = 3.71, p < .05$. This replication was therefore highly successful. It is interesting that here, within subjects, the preference reversal was even more pronounced than in Tversky and Kahneman's (1986) original between-subjects study. They found that 100% of participants chose dominating Option *B* and in their other group 58% preferred dominated Option *C*. Here, 91% chose dominating Options *B* but then 76% chose *C*.

Between subjects it is not accurate, however, to say that anyone actually "switched preferences," since no one is actually asked what their preferences are in both conditions. Conducting this study within subjects therefore provides us with the fascinating data of how many people actually switched preferences between conditions. Here, 71% of the sample (67 participants) preferred the dominating Option *B* in the first condition and then chose the

dominated Option C in the other. The 95% confidence interval for this 71% is .57 to .84, indicating a statistically significant within-subjects preference reversal.

The Princeton applicant

In Bastardi and Shafir's (1998) original between-subjects study, in the "accept vs. reject" condition, when there was no possibility of an A average, 43% of participants said they would reject while a majority, 57%, said they would accept the applicant. In the "accept, reject vs. wait" condition, when the grade average was either B or A, only 21% accepted the applicant, 5% rejected and a full 74% chose to wait. Those who chose to wait were then told that the real high school average was B and were then again asked whether they would accept or reject the applicant. Twenty-five percent accepted the applicant and 49% rejected.

Though Bastardi and Shafir (1998) discussed the possibility of a context effect in their between-subjects experiment, it was not realized until after this replication had been done that their between-subjects experiment and a within-subjects replication would not test the same hypotheses. The present replication cannot fulfill its original purpose then, which was to provide additional information regarding the within-subjects replication of between-subjects designs. So why include it? It was thought important to include these results regardless, for two reasons. First, it provides us with another interesting example of an instance where between- and within-subjects designs do not test the same hypotheses. Second, when this experiment is conducted within subjects, disjunction effects can be tested for. Disjunction effects cannot be tested for between subjects (Lambdin & Burdsal, 2007). This can provide us with valuable evidence given that it was argued not that long ago that disjunction effects do not even exist (Kuhberger, Komunska & Perner, 2001). Here are the results of the within-subjects replication:

Within-subjects replication data, the Princeton applicant (N = 95)

Participants who already accepted the applicant when the high school average was B should not choose to wait when the average is either B or A. Doing so is a violation of Savage's (1954) STP, a disjunction effect; it is to wait for information that should be obviously noninstrumental to the decision at hand.¹⁰ Here, of the 83% (79 participants) who accepted the applicant with a high school average of B, 41% (32 participants) then chose to wait for irrelevant information. The 95% confidence interval for this 41% is .26 to .55, indicating that the proportion of responses in which a disjunction effect was present was statistically significant. These results, therefore, provide additional evidence for the reality (and prevalence) of disjunction effects.

The child custody case

Recall that Parent A will now always represent the "impoverished" option and Parent B the "enriched" option. Shafir (1993) found that in the "award" group 36% of participants awarded custody to Parent A (64% to B) and that in the "deny" group 45% denied custody to A (55% to B). Below are the present replication results for the child custody case experiment.

Within-subjects replication data, the child custody case (N = 95)

Award A	Award B	Deny A	Deny B
65% (62)	35% (33)	34% (32)	66% (63)

¹⁰ This is not true for those who initially rejected the applicant with an average of B. For them, an average of B was perhaps not good enough and so when there is a possibility that the average is in fact A, waiting for that information is not irrational.

Figure 10. Within-subjects replication data, the child custody case

The overwhelming majority of participants wanted the impoverished option (Parent *A*) to have custody of the child. The enriched option (Parent *B*) was not both awarded and denied custody more than the impoverished option (Parent *A*) as hypothesized by Shafir (1993). Thus, the hypothesis offered by the theory of reason-based choice was not supported. Within subjects we can look at how many participants actually awarded and denied custody to the same parent. Out of 95 participants, only 3 (3%) did. With a 95% confidence interval of 0 to .08, this 3% is not significantly different from zero. As in Experiment 1, Shafir's (1993) finding failed to replicate.

The Asian disease problem

Tversky and Kahneman (1981) found that in the survival frame 72% of participants preferred Program *A* (28% *B*) and that in the mortality frame 78% preferred *D* (22% *C*). Here are the present within-subjects replication results:

Within-subjects replication data, the Asian disease problem (N = 95)

Program *A* Program *B* Program *C* Program *D*

63% (60) 37% (35) 33% (31) 67% (64)

Figure 11. Within-subjects replication data, the Asian disease problem

These results are again in line with Tversky and Kahneman's (1981) original findings. Recall that they found, between subjects, that in the survival frame the majority of participants preferred the sure thing whereas in the mortality frame the majority of participants opted for the gamble. The same result was here obtained within subjects. The 67% who preferred Program *D* over *C* was significantly different from the 50% that would be expected if there was no preference between programs, $z = 2.38, p < .05$. Though the 63% who preferred Program *A* over *B* was not significantly different from 50%, $z = 1.81, p > .05$, it is nevertheless the case that the overall qualitative pattern of data is the same as Tversky and Kahneman's (1981) original between-subjects study. Within subjects, 36% of participants (34 participants) chose Programs *A* and *D*. The 95% confidence interval for this 36% is .22 to .50, indicating a statistically significant preference reversal within subjects. (Only 3% of the sample chose Programs *B* and *C*.)

Second survey (N = 89)

Testing for between- vs. within-subjects differences using the same data

Above, the within-subjects results of the first survey were compared to the original between-subjects experiments replicated to see if findings held when switching from the one design to the other. A more direct way to test for this is to analyze the same data both within and between subjects. This is done by counterbalancing the order in which the conditions of each experiment are presented. Analyzing all participants' responses provides within-subjects data. Analyzing only the data for the first condition presented in each group provides between-subjects data. This procedure cannot, however, be performed on Bastardi and Shafir's (1998) Princeton applicant experiment, as for that experiment to work the conditions have to be presented in a particular order. Thus, what follows are Tversky and Kahneman's (1986) marbles lotteries, Shafir's (1993) child custody case and Tversky and Kahneman's (1981) Asian disease problem experiments counterbalanced so that the data can be analyzed both between and within subjects.

For the within-subjects data, there is 84% power to detect differences of 20 percentage points. For the between-subjects data, there is 54% power to detect such differences, since analyzing the data between subjects halves the sample size. Again, however, what we are primarily interested in is not achieving statistical significance but seeing whether the overall pattern of data is the same as the study being replicated. Here are the results for the marbles lotteries experiment:

The marbles lotteries, within (N = 89) vs. between subjects (n = 45 for A & B and 44 for C & D)

	Option A	Option B	Option C	Option D
Within subjects:	4% (4)	96% (85)	67% (60)	33% (29)
Between subjects:	4% (2)	96% (43)	61% (27)	39% (17)

Figure 12. The marbles lotteries, within vs. between subjects

For the marbles lotteries there is virtually no difference between the within- and between-subjects data. Actually, between subjects, participants correctly chose the dominating option by six more percentage points than within (see results for Options *C* and *D*). Within subjects, the 96% who preferred dominating Option *B* was significantly greater than 50%, $z = 6.91$, $p < .05$. The 67% who preferred dominated Option *C* over *D* was also significantly different from 50%, $z = 2.30$, $p < .05$. Between subjects, the 96% who preferred *B* was significantly greater than 50%, $z = 4.91$, $p < .05$. Due the halved sample size the 61% who preferred *C* was not significantly different 50%, $z = 1.05$, $p > .05$, though, again, the overall qualitative pattern of data was the same. The percentage of participants who switched preferences cannot be compared within and between subjects, since between subjects no one switches preferences.

The child custody case, within (N = 89) vs. between subjects (n = 45 for “award” and 44 for “deny”)

	Award A	Award B	Deny A	Deny B
Within subjects:	74% (66)	26% (23)	24% (21)	76% (68)

Between subjects: 73% (33) 27% (12) 23% (10) 77% (34)

Figure 13. The child custody case, within vs. between subjects

For the child custody case there is virtually zero difference between the within- and between subjects results. The vast majority of participants preferred Parent A to have custody to the child. Given that this result was also obtained between subjects, it is beginning to appear that Shafir's (1993) results are not replicable. This reinforces the idea suggested by the qualitative results of Experiment 1 that Shafir's (1993) experiment may lack construct validity. If Shafir's (1993) hypothesis had been supported, the percentage of participants who awarded and denied custody to Parent B would have both been significantly greater than 50%. Given that the data is not even in this direction, no statistical tests are needed. The percentage of participants who awarded and denied custody to the same parent cannot be compared, since between subjects no one is asked both who they will award and who they will deny custody to.

The Asian disease problem, within (N = 89) vs. between subjects (n = 45 for A & B and 44 for C & D)

Program A Program B Program C Program D

Within subjects:	62% (55)	38% (34)	34% (30)	66% (59)
Between subjects:	76% (34)	24% (11)	27% (12)	73% (32)

Figure 14. The Asian disease problem, within vs. between subjects

For the Asian disease problem, both the within- and between-subjects analyses of the data present a replication of Tversky and Kahneman's (1981) original result. Again, the percentage of participants who switched preferences cannot be compared since, between subjects, no one switches preferences. Within subjects, 62% preferred Program A, which is not significantly different from 50%, $z = 1.61$, $p > .05$. Sixty-six percent preferred Program D, which is significantly different from 50%, $z = 2.16$, $p < .05$. Between subjects, 76% preferred Program A and 73% preferred D. Both were significantly different from 50%, $z = 2.55$, $p < .05$ and $z = 2.22$, $p < .05$, respectively.

Discussion

When there are multiple options to choose from in each condition of an experiment, order effects can be tested for by either counterbalancing the order that conditions are presented in or by counterbalancing the order that the options within each condition are presented in (or both if

necessary). In Experiment 2, the first survey counterbalanced the order of options within each condition and the second survey counterbalanced the order of conditions. The former was done to address a concern that a reviewer had with a result from Experiment 1. Namely, in the Experiment 1 replication of Shafir's (1993) child custody case experiment, a reviewer thought that a majority of participants chose Parent *A* in both conditions simply because s/he was always the first option presented. This was not the case. Looking at the results of the first survey of Experiment 2, when the impoverished option was Parent *A*, 70% of the sample (33/47) awarded custody to the impoverished option. When the impoverished option was Parent *B*, 60% (29/48) awarded custody to the impoverished option. This difference was not significant, $z = 1.00$, $p > .05$, suggesting that the fact that a very small minority of participants awarded and denied custody to Parent *A* more than *B* in Experiment 1 was likely not caused by an order effect. Rather, it seems that the large percentage of participants who awarded and denied custody to the same parent (25%) in Experiment 1 was caused by the leading instructions used.

In the first survey there was 80% power to detect a 20 percentage-point difference in the data, to detect moderate to large differences in the data caused by order effects. No significant order effects were found and, if order effects were in fact present, none of them were of a magnitude large enough to cause a qualitative change in the overall pattern of data. To test for participants' sensitivity to differences along the "within-subjects continuum," in the first survey conditions were presented either together on the same page or separately on different pages. For all four replications this nonsignificantly enhanced the results. Again, there was 80% power to detect a 20 percentage-point difference in the data and for none of the replications did this effect alter the overall qualitative pattern of the results.

The second survey in Experiment 2 counterbalanced the order of conditions. This was done so that the data that resulted could be analyzed both within and between subjects. This provides a more direct comparison of the result of switching from one design to the other by removing any impact that using different samples might have on the data. Looking at the results of the second survey, all three within-subjects replications for which this comparison could be made showed nearly identical results both within and between subjects. As noted above, however, none of the attempts to replicate Shafir's (1993) child custody case experiment were met with success. The overwhelming majority of participants awarded custody to Parent *A* and denied custody to Parent *B*. The fact that most participants in Experiment 2 preferred Parent *A* to have custody even when tested for between subjects suggests that Shafir's (1993) own results may not be replicable.

Though the Princeton applicant replication's data cannot be compared to the between-subjects experiment on which it was based, it does nevertheless provide additional evidence for the existence of disjunction effects. Lambdin and Burdsal (2007) argued that disjunction effects can only be tested for within subjects because: 1) a disjunction effect is a violation of Savage's STP and 2) participants must always meet certain criteria before they can be said to violate the STP. Here, they must first accept the applicant when the high school average is reported as *B*. Those who do not accept the applicant at this time are not capable of violating the STP and so are no longer relevant to the experiment. In order to screen for this participant pool, a within-subjects design is necessitated. Here, of those who already accepted the applicant with a *B* average, 41% then went on to wait for what should have been obviously irrelevant information. (It does not matter if there is a possibility of an *A* average if a *B* average was already good enough.)

Interestingly, Bastardi and Shafir (1998) did not seem to pick up on the fact that a within-subjects replication does not test the same hypothesis as their between-subjects experiment. They stated, for instance, that the purpose of their own within-subjects replication was to test if the findings hold when the task is made transparent. But since the context effect is not present in the within-subjects variation, to compare the two would be like comparing apples and oranges. Further, their assumption that the within-subjects variation is transparent is not something they bothered to test. This assumption will here be put to test in Experiment 3.

CHAPTER 4

EXPERIMENT 3

Introduction

Experiment 2 explored whether between-subjects JDM experiments would successfully replicate within subjects. It also explored whether findings would change when the same data are analyzed both within and between subjects. Experiment 3 will more directly address the issue of transparency. It will do so by having another group of participants work through the same within-subjects replications used in Experiment 2, only here they will explicitly be asked to try and identify the research hypothesis for each. The child custody case (Shafir, 1993), Asian disease problem (Kahneman & Tversky, 1981) and Princeton applicant (Bastardi & Shafir, 1998) experiments were chosen because it was thought that many researchers would assume that each would be highly transparent within subjects. In fact, in Bastardi and Shafir's (1998) paper, the claim is made that a replication of a similar experiment yielded the results it did because it was "transparent." Here this assumption will be put to the test. The marbles lotteries (Kahneman & Tversky, 1986) was chosen as an example of an experiment that would likely remain largely "opaque" (not transparent) within subjects. It was hypothesized that the marbles lotteries experiment would prove to be the least transparent when replicated within subjects and that the child custody case experiment would prove to be the most transparent. In Experiment 3, a manipulation check will be employed to ensure that participants are in fact capable of correctly identifying the research hypothesis for an experiment. Participants will also be asked to rate their confidence that they are correct and to rank each scenario's transparency. This will tell us whether participants are overconfident and whether they have accurate insight into how transparent each design really was to them. Transparency was divided into three categories:

“completely transparent,” “somewhat transparent/somewhat opaque” and “not transparent at all/completely opaque.” In order to compare these transparency ratings to the proportions of participants who correctly identified the prediction for each replication, the following operational definitions were used: A replication would be classified as “completely transparent” if the proportion of participants who could correctly identify the prediction was significantly greater than 50%, “not transparent” if it was significantly lower than 50% and “somewhat transparent” if it was not significantly different from 50%. By having a separate sample perform these tasks in addition to answering all questions asked in Experiment 2, Experiment 3 will also address the possibility that asking participants to try and guess the prediction for each experiment they participate in might in some way alter their responses.

It was hypothesized that a majority of participants would correctly identify the prediction for the manipulation check but that a minority of participants would be able to for all four within-subjects replications. The reason for this hypothesis was the author’s contention that the typical JDM experiment is far less transparent, even when presented within subjects, than some researchers seem to assume (e.g., Bastardi & Shafir, 1998; Fischhoff, Slovic & Lichtenstein, 1979; Kahneman & Frederick, 2005). It was also hypothesized that though participants would, for the most part, be unable to discern the research hypothesis in question, that they would nevertheless be confident that they were correct. Finally, it was hypothesized that participants would not be able to accurately diagnose the transparency of the designs.

Method

Participants

Eighty undergraduate students at Wichita State University volunteered to participate for extra credit. Eight participants were dropped since they did not answer all of the questions, resulting in a sample size of 72.

Procedure

Experiment 3 had participants work through each experiment replicated in Experiment 2 (the Asian disease problem from Tversky & Kahneman, 1981; the marbles lotteries experiment from Tversky & Kahneman, 1986; the Princeton applicant experiment from Bastardi & Shafir, 1998; and the child custody case experiment from Shafir, 1993). After answering the questions for each of the four replications, participants were asked to guess the research hypothesis (or prediction) for that experiment. They were then asked to rate their confidence that they had guessed correctly (on a seven-point scale, three = “extremely confident,” negative three = “extremely unconfident”—this was then recoded into a seven-point scale with one = “extremely confident” and seven = “extremely unconfident”). They were also asked to rate the transparency of the design (on a three-point scale, one = “completely transparent,” two = “somewhat transparent” and three = “not transparent at all”). These concepts were presented to participants in the following text, which appeared on the first page of every questionnaire:

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.

The first scenario used served as a manipulation check. It was hypothesized that for all four replication scenarios used, a minority of participants would accurately identify the research hypothesis in question. The purpose of the manipulation check was therefore to ensure that participants can in fact accurately guess the research hypothesis of a within-subjects experimental scenario. The scenario that was used as the manipulation check was chosen because it was thought that it should be especially obvious what the research hypothesis was. Thus, if the instructions for Experiment 3 were clear to participants, they should be able to discern the prediction for the manipulation check. The scenario used as the manipulation check was adapted from Cialdini (1984/1993) and was as follows:

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.

As with Experiment 2, half of the participants saw both conditions of each experiment on the same page and half of the participants saw the conditions on different pages. This was done to see if separating the conditions affected participants’ accuracy, confidence and transparency ratings.

When coding what did and did not count as a correct identification of the research hypothesis the coder was as lenient as possible. This was intentionally done to avoid any claim that the results would have been different if only the coding was less strict. To count as a correct identification for the child custody case experiment, for example, participants only had to write something to the effect that changing the wording from award to deny would in some way impact the results. Thus, simply exhibiting some insight into what the manipulation was counted as successfully completing the task.

Materials

All data were collected using paper-and-pencil questionnaires. The instructions and questions used for Experiment 3 can be seen in Appendix B.

Results

The manipulation check

Nine out of 72 participants did not pass the manipulation check. In this section, their data will be compared to those who passed. A participant passed the manipulation check by accurately guessing the research hypothesis or prediction in question. Those who accurately guessed the prediction in question will be referred to as the “ST” (saw through) group and those who did not as the “DNST” (did not see through) group. All participants attempted to figure out the prediction for all within-subjects replications, as well as the manipulation check. For the nine that did not pass the manipulation check (DNST), there were 36 more opportunities to accurately identify a prediction (9 x 4: nine participants and the four other scenarios). For the DNST group, the research hypothesis for the four other scenarios was correctly identified twice out of 36 opportunities, or, 6% of the time (2/36). The ST group identified the research hypothesis 29 times out of 252 opportunities (63 x 4), or 12% of the time (29/252). The latter proportion (12%) is significantly different from zero (with a 95% confidence interval of .08 to .15) while the former is not (with a 95% confidence interval of -.02 to .13), though the 12% and 6% are not significantly different from each other, $z = 1.07$, $p > .05$. Below is a comparison of the transparency ratings for the ST and DNST groups.

Transparency ratings

To compare the transparency ratings for the DNST and ST groups, a chi square will be run for each of the five scenarios. Achieving a power of .8 would require a sample size of 108. Not much can be concluded from any statistical analyses of these groups, given that there are only nine participants in the ST group. .

The manipulation check transparency ratings, DNST (n = 9) vs. ST (n = 63)

	Completely	Somewhat	Not transparent
DNST:	11% (1)	78% (7)	11% (1)
ST:	40% (25)	57% (36)	3% (2)

Figure 15. The manipulation check transparency ratings, DNST vs. ST

The transparency ratings somewhat reflect the difference between the two groups, with 40% of the ST group and only 11% of the DNST group calling that scenario “completely transparent.” A chi square was run, however, and was not significant, $\chi^2 (2, N = 72) = 3.53, p > .05$.

The marbles lotteries transparency ratings, DNST (n = 9) vs. ST (n = 63)

	Completely	Somewhat	Not transparent
DNST:	11% (1)	67% (6)	22% (2)
ST:	17% (11)	51% (32)	32% (20)

Figure 16. The marbles lotteries transparency ratings, DNST vs. ST

With the marbles lotteries there is not much of a difference in the transparency ratings between groups. A chi square was run and was not significant, $\chi^2 (2, N = 72) = .80, p > .05$.

The Princeton applicant transparency ratings, DNST (n = 9) vs. ST (n = 63)

	Completely	Somewhat	Not transparent
DNST:	11% (1)	56% (5)	33% (3)
ST:	14% (9)	60% (38)	25% (16)

Figure 17. The Princeton applicant transparency ratings, DNST vs. ST

Again, there is no significant difference between the groups, $\chi^2 (2, N = 72) = .30, p > .05$.

The child custody case transparency ratings, DNST (n = 9) vs. ST (n = 63)

	Completely	Somewhat	Not transparent
DNST:	22% (2)	78% (7)	0% (0)
ST:	25% (16)	48% (30)	27% (17)

Figure 18. The child custody case transparency ratings, DNST vs. ST

Here, there is more of a difference with the “somewhat transparent” and “not transparent” categories, but the chi square was not significant, $\chi^2 (2, N = 72) = 3.85, p > .05$.

The Asian disease problem transparency ratings, DNST (n = 9) vs. ST (n = 63)

	Completely	Somewhat	Not transparent
DNST:	11% (1)	78% (7)	11% (1)
ST:	27% (17)	57% (36)	16% (10)

Figure 19. The Asian disease problem transparency ratings, DNST vs. ST

Here too there is no significant difference between the groups, $\chi^2(2, N = 72) = 1.47, p > .05$. It can be seen, therefore, that the transparency ratings for the DNST and ST groups do not reliably distinguish one group from the other.

All scenarios' confidence ratings, DNST (n = 9) vs. ST (n = 63)

For the confidence ratings, one = “extremely confident” and seven = “extremely unconfident.” The DNST participants were, for all but one scenario, slightly less confident in their accuracy. This difference was only significant for the manipulation check. The following table contains the means and standard deviations for both group’s confidence ratings. It also contains a *t* test for each comparison. The power for the *t* tests is .28.

TABLE 1
AVERAGE CONFIDENCE RATINGS, DNST VS. ST

Experiment name	DNST	ST	<i>t</i> test
Manipulation check	3.11 (1.45)	2.21 (1.03)	$t(70) = 2.33, p < .05$
Marbles lotteries	4.67 (1.66)	3.59 (1.61)	$t(70) = 1.87, p > .05$
Princeton applicant	4.33 (1.94)	3.75 (1.53)	$t(70) = 1.04, p > .05$
Child custody case	3.22 (1.86)	3.33 (1.41)	$t(70) = -.21, p > .05$

Asian disease problem	3.44 (1.88)	2.95 (1.42)	$t(70) = .93, p > .05$
-----------------------	-------------	-------------	------------------------

Replication data

Below are comparisons of the replication data for the DNST and ST groups. Due to the reduced sample sizes, there is 63% power to detect differences of 20 percentage points for the ST group and only 13% power to detect such differences in the DNST group.

The marbles lotteries data, DNST (n = 9) vs. ST (n = 63)

	Option A	Option B	Option C	Option D
DNST:	0% (0)	100% (9)	89% (8)	11% (1)
ST:	5% (3)	95% (60)	59% (37)	41% (26)

Figure 20. The marbles lotteries data, DNST vs. ST

In the DNST group, 100% preferred dominating Option *B* in the first condition and then 89% preferred dominated Option *C* in the second. The 100% was significantly different from the 50% that would be expected if there was no preference between options, $z = 2.45, p < .05$. Due to the small sample size ($n = 9$), the 89% was not significantly different from 50%, $z = 1.80, p > .05$. In the ST group, in the first condition 95% of participants preferred dominating Option *B*

and in the second 59% preferred dominated Option C. The 95% was significantly different from 50%, $z = 5.66$, $p < .05$, but the 59% was not, $z = 1.01$, $p > .05$. In both cases, however, what really matters is that the overall qualitative pattern of results was the same, with the majority preferring the dominating option in the first condition and the dominated option in the second. Thus, the experiment replicated successfully for both groups.

The Princeton applicant data, DNST (n = 9) vs. (n = 63)

In the DNST group, of the seven participants (78%) who already accepted the applicant with a high school average of B, 57% (four) then chose to wait for noninstrumental information when the high school average was either B or A. In the ST group, of the 54 (86%) who had already accepted with a B average, 56% (30) then opted to wait for irrelevant information. If you already accepted the applicant with a high school average of B, then you should also accept the applicant when the high school average is either B or A. The 95% confidence interval for the proportion of participants who violated the STP in the DNST group is .43 to .71. For the ST group it is .41 to .70. In both groups a disjunction effect was present in a statistically significant proportion of participants' decisions.

The child custody case data, DNST (n = 9) vs. (n = 63)

	Award A	Award B	Deny A	Deny B
DNST:	100% (9)	0% (0)	11% (1)	89% (8)
ST:	67% (42)	33% (21)	30% (19)	70% (44)

Figure 21. The child custody case data, DNST vs. ST

In the DNST group, 100% awarded custody to Parent A. This was significantly different from the 50% that would be expected if there was no preference, $z = 2.45, p < .05$. Also in the DNST group, 89% denied custody to Parent B. This was not significantly different from 50%, $z = 1.80, p > .05$. For the ST group, 67% awarded custody to Parent A. This was not significantly different from 50%, $z = 1.94, p > .05$. Seventy percent denied custody to Parent B. This was significantly different from 50%, $z = 2.29, p > .05$. Again, what is crucial is that the same results are found in both groups. Both groups preferred Parent A to have custody of the child in both conditions, and, as a result, the child custody case experiment failed to replicate for both groups.

The Asian disease problem data, DNST (n = 9) vs. ST (n = 63)

	Program A	Program B	Program C	Program D
DNST:	78% (7)	22% (2)	44% (4)	56% (5)
ST:	62% (39)	38% (24)	40% (25)	60% (38)

Figure 22. The Asian disease problem data, DNST vs. ST

In the DNST group, 78% of participants preferred Program A in the first condition while 56% preferred D in the second. Neither were significantly different from 50%, $z = 1.24, p > .05$ and $z = .26, p > .05$, respectively. In the ST group, 62% preferred Program A in the first condition and 60% preferred D in the second. Again, neither were significantly different from 50%, $z = 1.36, p > .05$ and $z = 1.13, p > .05$, respectively. In both groups the majority of participants preferred Programs A and D, indicating that the Asian disease problem successfully replicated for both groups.

The purpose that the manipulation check served was to ensure that participants can in fact accurately guess the research hypothesis of a within-subjects experimental scenario. The manipulation check fulfilled this purpose, with a full 88% of the sample (63 participants) correctly identifying the prediction for that particular scenario. Since none of the findings changed between the DNST and ST groups (or at least, since there were so few people in the ST group), we can safely collapse across them.

Different vs. same page

In Experiment 3, as in Experiment 2, whether the conditions for each scenario were presented on different pages vs. than on the same page was varied.

All scenarios' proportions that correctly identified prediction, different (n = 34) vs. same (n = 38) page

Thirty-four participants saw the conditions on different pages and 38 on the same page. The following table contains the proportions of participants who correctly identified the research hypothesis for each scenario for both the same-page and different-pages groups. Though here there is only 40% power to detect 20 percentage-point differences, it can easily be seen that in every case the proportions are almost identical.

TABLE 2

PERCENT IDENTIFIED PREDICTION, DIFFERENT VS. SAME PAGE

Experiment name	Same page	Different pages	z test
Manipulation check	89% (34/38)	85% (29/34)	$z = .51, p > .05$
Marbles lotteries	5% (2/38)	0% (0)	$z = 1.32, p > .05$
Princeton applicant	3% (1/38)	0% (0)	$z = 1.02, p > .05$
Child custody case	8% (3/38)	6% (2/34)	$z = .33, p > .05$
Asian disease problem	32% (12/38)	32% (11/34)	$z = .00, p > .05$

Below are comparisons of the transparency ratings for the same- vs. different-pages groups.

Transparency ratings

To compare the transparency ratings for the same- vs. different-pages groups, a chi square was calculated for each experiment. The chi squares have a power of .62.

The manipulation check, transparency ratings for different (n = 34) vs. same (n = 38) page

	Completely	Somewhat	Not transparent
Same page:	45% (17)	55% (21)	0% (0)
Different pages:	26% (9)	65% (22)	9% (3)

Figure 23. The manipulation check, transparency ratings for different vs. same page

The largest difference is between the proportions of participants who rated the scenario “completely transparent” when the conditions were presented on the same page vs. when they were presented on different pages. The chi square was, however, nonsignificant, $\chi^2(2, N = 72) = 5.28, p > .05$.

The marbles lotteries, transparency ratings for different (n = 34) vs. same (n = 38) page

	Completely	Somewhat	Not transparent
Same page:	24% (9)	47% (18)	29% (11)
Different pages:	9% (3)	59% (20)	32% (11)

Figure 24. The marbles lotteries, transparency ratings for different vs. same page

Again, the chi square was not significant, $\chi^2 (2, N = 72) = 2.89, p > .05$.

The Princeton applicant, transparency ratings for different (n = 34) vs. same (n = 38) page

	Completely	Somewhat	Not transparent
Same page:	13% (5)	58% (22)	29% (11)
Different pages:	15% (5)	62% (21)	24% (8)

Figure 25. The Princeton applicant, transparency ratings for different vs. same page

Here, the results are almost identical, $\chi^2 (2, N = 72) = .28, p > .05$.

The child custody case, transparency ratings for different (n = 34) vs. same (n = 38) page

	Completely	Somewhat	Not transparent
Same page:	29% (11)	50% (19)	21% (8)
Different pages:	21% (7)	53% (18)	26% (9)

Figure 26. The child custody case, transparency ratings for different vs. same page

Again, the results are nearly identical, $\chi^2 (2, N = 72) = .75, p > .05$.

The Asian disease problem, transparency ratings for different (n = 34) vs. same (n = 38) page

	Completely	Somewhat	Not transparent
Same page:	26% (10)	63% (24)	11% (4)
Different pages:	21% (7)	56% (19)	24% (8)

Figure 27. The Asian disease problem, transparency ratings for different vs. same page

For the Asian disease problem, $\chi^2(2, N = 72) = 2.23, p > .05$. For none of the replications did presenting conditions on the same vs. different pages result in a significant difference in the transparency ratings.

All scenarios' confidence ratings, different (n = 34) vs. same (n = 38) page

The following table contains the average confidence ratings and their standard deviations for each experiment, for both the same-page and different-pages groups. It also contains a t test for each comparison. The t tests have a power of .55. For the confidence ratings, one = “extremely confident” and seven = “extremely unconfident.”

TABLE 3

AVERAGE CONFIDENCE RATINGS, DIFFERENT VS. SAME PAGE

Experiment name	Same page	Different pages	t test
Manipulation check	1.92 (.94)	2.76 (1.16)	$t(70) = 3.41, p < .05$
Marbles lotteries	3.37 (1.75)	4.12 (1.45)	$t(70) = 1.97, p > .05$
Princeton applicant	3.53 (1.66)	4.15 (1.46)	$t(70) = 1.68, p > .05$
Child custody case	3.03 (1.38)	3.65 (1.50)	$t(70) = 1.83, p > .05$
Asian disease problem	2.79 (1.42)	3.26 (1.52)	$t(70) = 1.37, p > .05$

The same- vs. different-pages manipulation did not have a significant (or consistent) impact on any of the transparency ratings. Looking at the average confidence ratings, a significant difference was only found with the manipulation check. It is therefore safe to collapse across this manipulation. The replication data will not here be analyzed for different- vs. same-page differences as such an analysis was already performed in Experiment 2. Recall that it was found that, just as it was with the transparency ratings above, for all four replications the manipulation nonsignificantly impacted the results. Below is a look at the entire sample, collapsing across the same- vs. different-pages and DNST vs. ST groups.

All data combined

All scenarios' proportions that correctly identified predictions (N = 72)

The following table contains the proportions of participants who correctly discerned the research hypothesis for each experiment, as well as that proportion's 95% confidence interval.

TABLE 4
PERCENT IDENTIFIED PREDICTION

Experiment name	% (#)	95% CI
Manipulation check	88% (63)	.78 to .97
Marbles lotteries	3% (2)	0 to .07
Princeton applicant	1% (1)	0 to .04
Child custody case	7% (5)	.01 to .13
Asian disease problem	32% (23)	.21 to .43

Excluding the manipulation check and the Asian disease problem, only five participants correctly identified a research hypothesis for any of the scenarios. Again excluding the manipulation check and the Asian disease problem, no participant who identified the research hypothesis for one scenario also identified it for either of the other two. Four out of five of these

participants also identified the research hypothesis for the Asian disease problem. Only for the manipulation check did a majority of participants correctly discern the research hypothesis. With the Asian disease problem, though the proportion of participants who correctly identified the research hypothesis was significantly different from zero, still only a minority of participants was able to do so; and, this number was arrived at by including everyone who exhibited any understanding of what the manipulation was. This was done with all of the within-subjects replications, in order to be as liberal as possible. For instance, with the marbles lotteries replication, no one actually correctly identified the prediction, the 3% cited above consisted of people who simply stated what the manipulation was. Almost all guesses concerning what was being predicted for each replication had absolutely nothing to do with the manipulation. For example, for Shafir's (1993) child custody case, almost everyone wrote that they wanted what was best for the child.

What follows is an analysis of the transparency ratings for the entire sample.

Transparency ratings

The manipulation check transparency ratings (N = 72)

Completely	Somewhat	Not transparent
36% (26)	60% (43)	4% (3)

Figure 28. The manipulation check transparency ratings

The majority of participants thought that the manipulation check was “somewhat transparent.” Since 88% of participants were able to correctly identify the prediction, a rating of “completely transparent” would have been warranted according to the present operational definition (being significantly greater than 50%; here, the 95% confidence interval is .78 to .97).

The marbles lotteries transparency ratings (N = 72)

Completely	Somewhat	Not transparent
17% (12)	53% (38)	31% (22)

Figure 29. The marbles lotteries transparency ratings

The majority of participants thought that the marbles lotteries replication would be “somewhat transparent.” Since the proportion of participants who could correctly identify the prediction was not significantly different from zero (with the 95% confidence interval being 0 to .07), this rating is inaccurate. Actually, no one correctly stated what the prediction was. The two participants who were counted as successfully completing the task simply stated what the manipulation was.

The Princeton applicant transparency ratings (N = 72)

Completely	Somewhat	Not transparent
14% (10)	60% (43)	26% (19)

Figure 30. The Princeton applicant transparency ratings

Here again, the majority of participants chose “somewhat transparent.” This rating is inaccurate given that only 1% of the sample was accurate (with a 95% confidence interval of 0 to .04). Almost everyone wrote something akin to wanting to give the applicant a chance or wanting to find the best applicants possible.

The child custody case transparency ratings (N = 72)

Completely	Somewhat	Not transparent
25% (18)	51% (37)	24% (17)

Figure 31. The child custody case transparency ratings

Here, the 95% confidence interval for the proportion of participants who successfully completed the task was .01 to .13. Thus, the majority rating of “somewhat transparent” should be considered inaccurate.

The Asian disease problem transparency ratings (N = 72)

Completely Somewhat Not transparent

25% (18) 60% (43) 15% (11)

Figure 32. The Asian disease problem transparency ratings

Sixty percent of participants thought that the Asian disease problem, presented within subjects, would be “somewhat transparent.” It is debatable whether this is an accurate rating or not. According to the present operational definition of “somewhat transparent” (being significantly lower than 50%) this rating would be considered inaccurate. Thirty-two percent of participants successfully completed the task (with a 95% confidence interval of .21 to .43). The author realizes that this present operational definition is somewhat arbitrary and that many would likely consider 32% large enough to fully warrant the label “somewhat transparent.” If we agree with this, then participants accurately identified the transparency of the within-subjects replications one out of four times (25% success rate).

Looking at the graphs above it can be seen that a majority of participants thought that every scenario was “somewhat transparent.” Speculating why is difficult as we do not know what exactly was going on in the heads of participants. Did participants tend to choose “somewhat transparent” simply because they considered it the “safest” choice? Were they thinking that their own guess was likely “somewhat transparent” or that for everyone on average it would be “somewhat transparent”? Future research should be more explicit in attempting to identify this.

All scenarios’ average confidence ratings (N = 72)

The following table contains the average confidence rating and its standard deviation for each experiment, as well as its 95% confidence interval. For the confidence ratings, one = “extremely confident” and seven = “extremely unconfident.”

TABLE 5
AVERAGE CONFIDENCE RATINGS

Experiment name	M (SD)	95% CI
Manipulation check	2.32 (1.12)	2.06 to 2.58
Marbles lotteries	3.72 (1.65)	3.34 to 4.10
Princeton applicant	3.82 (1.59)	3.45 to 4.19
Child custody case	3.32 (1.46)	2.98 to 3.66
Asian disease problem	3.01 (1.48)	2.67 to 3.35

For the manipulation check, the average was closest to “very confident.” For the marbles lotteries and the Princeton applicant the average was closest to “neither confident nor unconfident” and for the child custody case and the Asian disease problem it was closest to “somewhat confident.” What follows is a comparison of the replication results of Experiments 2 and 3. The purpose of the comparison is to see if we can combine the two samples for a larger replication sample size.

Replication data, Experiment 2 vs. Experiment 3

Below is a comparison of the replication results from Experiments 2 and 3. For the Experiment 2 data there is 86% power to detect differences of 20 percentage points. For Experiment 3, there is 75% power to detect such differences.

The marbles lotteries, Experiment 2 (N = 95) vs. Experiment 3 (N = 72)

	Option A	Option B	Option C	Option D
Experiment 2 data:	9% (9)	91% (86)	76% (71)	24% (23)
Experiment 3 data:	4% (3)	96% (69)	63% (45)	38% (27)

Figure 33. The marbles lotteries, Experiment 2 vs. Experiment 3

In Experiment 2, for the marbles lotteries, in the first condition 91% of participants preferred dominating Option *B* and in the second condition 76% then preferred dominated Option *C*. Both of these proportions were significantly different from the 50% that would be expected if there was preference between options, $z = 6.20, p < .05$ and $z = 3.71, p < .05$, respectively. In Experiment 3, 96% of participants chose *B* and 63% chose *C*. The 96% was significantly different from 50% but the 63% was not, $z = 6.22, p < .05$ and $z = 1.57, p < .05$, respectively. Though the 63% was not significantly different from 50%, what is important is that it was still the case that a majority of participants then preferred the dominated option when the options were rewoded. Thus, both Experiment 2 and Experiment 3 yielded a successful within-subjects replication of Tversky and Kahneman's (1986) between-subjects finding.

The Princeton applicant, Experiment 2 (N = 95) vs. Experiment 3 (N = 72)

In Experiment 2, of the 89% (79) of participants who accepted the applicant with a high school average of B, 41% (32) then chose to wait for noninstrumental information. In Experiment 3, of the 85% (61 participants) who accepted the applicant with a B average, 56%

(34) then chose to wait for noninstrumental information. The 95% confidence interval for the 41% of participants who violated the STP in Experiment 2 is .26 to .55. For the 56% in Experiment 3 it is .41 to .70. In both Experiments 2 and 3 a disjunction effect was present in a significant proportion of participants' decisions.

The child custody case, Experiment 2 (N = 95) vs. Experiment 3 (N = 72)

	Award A	Award B	Deny A	Deny B
Experiment 2 data:	65% (62)	35% (33)	34% (32)	66% (63)
Experiment 3 data:	71% (51)	29% (21)	28% (20)	72% (52)

Figure 34. The child custody case, Experiment 2 vs. Experiment 3

Here again the data from Experiments 2 and 3 follow exactly the same pattern. Neither provided a replication of Shafir (1993); and, again, both groups of participants greatly preferred Parent A to have custody of the child in both conditions. In neither group did participants award and deny custody to Parent B more than A. As stated above, it does not seem that Shafir's (1993) original finding is replicable.

The Asian disease problem, Experiment 2 (N = 95) vs. Experiment 3 (N = 72)

	Program A	Program B	Program C	Program D
Experiment 2 data:	63% (60)	35% (33)	34% (32)	66% (63)
Experiment 3 data:	64% (46)	36% (26)	40% (29)	60% (43)

Figure 35. The Asian disease problem, Experiment 2 vs. Experiment 3

Here too the same patterns of data were found between Experiments 2 and 3. In Experiment 2, 63% of participants preferred Program A in the first condition and 66% preferred D in the second. The 63% was not significantly different from 50%, while the 66% was, $z = 1.80, p > .05$ and $z = 2.23, p < .05$, respectively. In Experiment 3, 64% of participants preferred A in the first condition and 60% preferred D in the second. Neither proportion was significantly different from 50%, $z = 1.70, p > .05$ and $z = 1.21, p > .05$, respectively. In both Experiment 2 and Experiment 3, however, the same overall qualitative pattern of results was found, and both constituted successful replications of Tversky and Kahneman's (1981) results.

In both Experiments 2 and 3, the marbles lotteries and the Asian disease problem replicated successfully. Also in both, a significant disjunction effect was present in the within-subjects variation of the Princeton applicant. Similarly, the child custody case experiment did

not successfully replicate in either Experiment 2 or 3. Given that the exact same results were found in both Experiments, the present replication data from Experiment 3 can be combined with the data from Experiment 2, thereby yielding a larger sample size. Further, by having a separate sample try to identify the prediction for each replication in addition to answering all questions asked in Experiment 2, Experiment 3 allowed us to address the possibility that asking participants to try and guess the prediction for each experiment they participate in might in some way alter their responses. This was not the case.

Replication data, Experiments 2 & 3 combined

What follows is an analysis of the replication data from Experiments 2 and 3 combined. There is 99.9% power to detect differences of 20 percentage points.

The marbles lotteries data, Experiments 2 & 3 combined (N = 256)

Option A	Option B	Option C	Option D
6% (16)	94% (240)	70% (178)	30% (77)

Figure 36. The marbles lotteries data, Experiments 2 & 3 combined

Tversky and Kahneman (1986) found, between subjects, that 100% of participants preferred dominating Option *B* ($N = 88$) while 58% of participants chose dominated Option *C* ($N = 124$). Here, within subjects, 94% of participants chose dominating Option *B* while 70% preferred dominated Option *C*. Both proportions are significantly different from 50%, $z = 11.09$, $p < .05$ and $z = 4.62$, $p < .05$. This is a strong and successful replication. Within subjects we can look at how many people actually switched preferences between conditions. Here, within subjects, 66% of the sample (169 participants) chose Options *B* and *C*. The 95% confidence interval for that 66% is .52 to .80, indicating a statistically significant preference reversal.

The Princeton applicant data, Experiments 2 & 3 combined ($N = 256$)

Eighty-four percent of the sample (216 participants) accepted the applicant with a high school average of *B*. Of the 84% who already accepted the applicant with a *B* average, 48% (104 participants) went on to wait for irrelevant information. The 95% confidence interval for that 48% is .34 to .63, indicating that a disjunction effect was present in a significant proportion of participants' decisions.

The child custody case data, Experiments 2 & 3 combined ($N = 256$)

Award <i>A</i>	Award <i>B</i>	Deny <i>A</i>	Deny <i>B</i>
70% (180)	30% (76)	28% (72)	72% (184)

Figure 37. The child custody case data, Experiments 2 & 3 combined

Recall that in Shafir's (1993) study the majority of participants awarded and denied custody to the enriched option, Parent *B*. Within subjects, only 4% of the sample awarded and denied custody to the same parent. The 95% confidence interval for that 4% is $-.02$ to $.10$. Thus, the proportion of participants who awarded and denied custody to the same parent was not significantly different from zero. The overwhelming majority of participants awarded the impoverished option and denied the enriched. Participants overwhelmingly chose Parent *A* to have custody of the child in both conditions. As noted above, it does not seem that Shafir's (1993) original finding is replicable. In Experiment 2, the original finding also failed to replicate between subjects.

The Asian disease problem data, Experiments 2 & 3 combined (N = 256)

Program A	Program B	Program C	Program D
63% (161)	37% (95)	35% (90)	65% (166)

Figure 38. The Asian disease problem data, Experiments 2 & 3 combined

Tversky and Kahneman (1981) found, between subjects, that 72% of participants preferred Program A, 28% *B* and that 22% preferred *C* and 78% *D*. Here, within subjects, 63% preferred A, 37% *B*, 35% *C* and 65% *D*. Thus, the finding was the same though slightly less dramatic. With the increased sample size, both majorities—the 63% and the 65%—were significantly different from 50%, $z = 2.97$, $p < .05$ and $z = 3.43$, $p < .05$, respectively. Running the experiment within subjects has the added advantage of providing data on how many participants actually switch preferences between frames or conditions. Such data speaks directly to the robustness of the framing effect. Here, 5% (12) chose Programs *B* and *C* and 33% (85) chose Programs A and *D*. The respective 95% confidence intervals for these proportions are 0 to .11 and .20 to .47. Thus, switching from Program *B* to *C* was not a significant preference reversal while switching from Program A to *D* was.

Discussion

For the four replication scenarios, those who passed the manipulation check identified the research hypothesis in question 12% of the time, whereas those who did not identified it 6% of

the time. Neither group, however, produced impressive results. Though distinguishing between those who did and did not pass the manipulation check did not really alter the data in any meaningful way, the manipulation check did fulfill its purpose. For all replication scenarios save the Asian disease problem (with 32% successfully completing the task), a stark minority of participants could identify the prediction when asked. For the manipulation check, however, 88% of participants correctly identified the prediction. The manipulation check therefore helped lend validity to the experiment by demonstrating that participants' poor performance on the within-subjects replications was not due to any universal inability of theirs to look at an experimental scenario and identify its research hypothesis.

Looking at the transparency ratings, it was hypothesized that the child custody case would be the most highly transparent when presented within subjects. This did not prove to be the case. Though the proportion of participants who correctly identified the prediction for the child custody case was significantly different from zero (7%, with a 95% confidence interval of .01 to .13), this is still only five out of 72 participants, hardly anyone. This was a surprising finding. Participants were most successful with the Asian disease problem, with 23 out of 72 getting it right, though here still well less than half of participants (32%) could identify the prediction in question. The Asian disease problem, then, could be classified as being "somewhat transparent." Perhaps this accounted for the fact that though it still replicated successfully, the finding was not as dramatic as Tversky and Kahneman's (1981) original. For the marbles lotteries, 3% (2) correctly identified the research hypothesis and for the Princeton applicant scenario, only one participant (1% of the sample) could do so. One could argue that this perhaps indicates that the Asian disease problem is in fact somewhat more transparent than the other experiment replicated within subjects. A problem with this conclusion is the possibility that a

number of participants may have already been familiar with Tversky and Kahneman's (1981) famous study. It is occasionally lectured on in the department and is also highly memorable.

It should be added again that these figures were produced by being very lenient with what counted as a successful identification of the research hypothesis. For example, in order to successfully guess the research hypothesis for the child custody case, participants only had to say something to the effect that changing the wording from award to deny would impact the results. And yet only five out of 72 participants guessed correctly. Instead, nearly everyone wrote something about studying what attributes make for a good parent. What this means is that simply stating what the manipulation was was accepted as correct. Kahneman and Frederick (2005) have argued that in a within-subjects design, when studying heuristics and judgments, participants automatically know what the manipulation is and that this is enough to confound the study. Clearly this is not the case. For both the marbles lotteries and the Asian disease problem, the majority of participants did not seem to automatically know what the manipulation was, and, whether they did or did not, both replicated just fine. The only replication that can be considered "somewhat transparent" is the Asian disease problem. Though the original between-subjects finding did successfully replicate, the results of the within-subjects replication were not as dramatic as the original results. It does not seem, therefore, that within-subjects transparency resulted in demand characteristics that in any way inflated the results.

Even though for all replication scenarios (not including the manipulation check) a minority of participants could identify the research hypothesis in question, a majority of participants rated every scenario as being "somewhat transparent." With the Princeton applicant scenario, for instance, 60% of participants thought that it was "somewhat transparent" (i.e., somewhat easy to guess the research hypothesis in question) and 14% "completely transparent,"

even though only one participant correctly identified the research hypothesis. With the marbles lotteries, 17% of the sample thought that it was “completely transparent” and 53% “somewhat transparent,” even though only two participants correctly identified the prediction. In order to have been even somewhat accurate, a majority of participants should have chosen “not transparent” for all four replication scenarios except perhaps the Asian disease problem.

This was similarly the case with the confidence ratings, where for all four replication scenarios participants were overconfident in their accuracy. For the Asian disease problem, participants were on average “somewhat confident” that they had accurately identified the research hypothesis for that scenario, even though only 32% actually did. For the child custody case it was worse. Participants were still “somewhat confident” they had correctly guessed the prediction when really only 7% were correct. For the Princeton applicant scenario and the marbles gambles participants were “neither confident nor unconfident” they were correct when only 1 and 3% were accurate. Since a minority of participants were correct in each instance, for all four replication scenarios the average response should have been somewhere in the “unconfident” camp. Again, the manipulation check displayed that participants were capable of being accurate. Here 88% of participants were correct and they were on average “very confident” that they were right. In all four of the replications, however, overconfidence was in evidence.

Combining the replication data from Experiments 2 and 3, it can be seen that the marbles lotteries and Asian disease problem experiments both successfully replicated within subjects. The child custody case experiment, however, did not. The way in which Shafir (1993) originally analyzed the results for the child custody case experiment was not appropriate for a between-subjects design, but this does not mean that his original hypothesis cannot be tested within

subjects. The prediction for the original between-subjects child custody case experiment was that the enriched option (Parent *B*) will be awarded and denied custody more than the impoverished option (Parent *A*). Shafir (1993) found that 64% awarded and 55% of participants denied custody to Parent *B*. Here, within subjects, only 30% of participants awarded custody to Parent *B* while 72% denied Parent *B* custody. Participants overwhelmingly wanted Parent *A* (the impoverished option) to have custody of the child, with 70% awarding and only 28% denying. Here, only 4% of participants (10/256) awarded and denied custody to the same parent, and this was not significantly different from zero (the 95% confidence interval being 0 to .10).

This study, therefore, did not replicate within subjects. Furthermore, in Experiment 2, Shafir's (1993) study also failed to replicate between subjects, where it was also found that the overwhelming majority of participants wanted Parent *A* to have custody in both conditions. This leads the author to conclude that the problem is likely not with the replications but with Shafir's (1993) original study. Even his original results, for instance, were not that impressive. Analyzed appropriately, neither the 64% who awarded *B* nor the 55% who denied *B* were significantly different from the 50% that would be expected if there was no preference between options, $z = 1.84, p > .05$ and $z = .65, p > .05$, respectively. In Experiment 1, 19% of participants commented (without solicitation) that Parent *B* was a more unfit parent simply because s/he had an active social life. This was interpreted as suggesting that this experiment might lack construct validity, since many participants do not seem to agree with Shafir's (1993) assumption that Parent *B* has both more extreme negative and extreme positive attributes than *A*. It is also an interesting possibility though that this result may be indicative of regional differences (as suggested by one reviewer who indelicately declared that s/he did not care what people at "Podunk University" in Kansas think. Even if this were the correct explanation for the abject failure of Shafir's (1993)

results to replicate, the fact would remain that at Podunk U the experiment likely lacks construct validity.)

In conclusion, participants were not accurate in their transparency ratings and were also highly overconfident regarding said accuracy. For half of the four replication scenarios, the proportion of participants who were actually able to discern the research hypothesis in question was not significantly different from zero. For the other two it was, but was still only a minority (7% and 32%). Thus, though participants believed the scenarios to be somewhat transparent, only the Asian disease problem could realistically be considered so. A success rate of 25% is not indicative of accuracy.

As discussed in the Introduction of this paper, the overall goal is to address two assumptions. One of these assumptions (that between- and within-subjects designs—issues of transparency aside—are interchangeable) was addressed at length in the Introduction. The other assumption was that between-subjects designs are not transparent while within-subjects designs are. Experiment 3 provides some evidence that replicating between-subjects experiments using a within-subjects design does not always render the task “transparent,” as sometimes assumed by researchers (e.g., Bastardi & Shafir, 1998; Fischhoff, Lichtenstein & Slovic, 1979; Kahneman & Frederick, 2005). This would further indicate that many within-subjects designs in the field may not be as transparent as often assumed. Another point made by Experiment 3 is that the transparency of any within-subjects design is an empirical question (and should be treated as such) and should never be assumed *a priori*. If the participants of Experiment 3 had simply been asked to work through each design and rate its transparency, the conclusion would have been made that each and every one of them was at least somewhat transparent. Looking at their actual

performance in attempting to discern the research hypothesis (or manipulation), however, this was not the case.

CHAPTER 5

EXPERIMENT 4

Introduction

The purpose of Experiment 4 is to test whether participants who are already familiar with the between-subjects experiments used in Experiments 1 through 3, as well as with experimental design in general, will be able to accurately assess how transparent within-subjects replications of each actually are. This was done by using members of the Society for Judgment and Decision Making (SJDM) as participants. It was hypothesized that, because of the hindsight bias, SJDM members would tend to overestimate how transparent the within-subjects variations are to typical psychology research participants.

Research on the hindsight bias demonstrates that once we have learned something, we are typically incapable of unlearning it (Fischhoff, 1975). This can be problematic when engaged in tasks that require us to suppress or ignore information that we have already encoded. SJDM members are highly familiar with research methodology in general and the overwhelming majority of them should already be familiar with the experiments replicated herein. Because of this, it was hypothesized that the hindsight bias would induce them to exaggerate how transparent these replications are perceived to be by participants who do not possess the knowledge they are privy to.

In Experiment 3, the transparency categories of “completely,” “somewhat” and “not transparent” were operationalized as the proportion of participants who were able to discern the prediction in question as being significantly greater than 50%, not significantly different from 50% and significantly lower than 50%, respectively. This operationalization was abandoned for the Asian disease problem. Here, in Experiment 3, 32% of participants “saw through” the

design. Since many would likely consider this proportion to count as being “somewhat transparent,” this interpretation will also be used when looking at the results of Experiment 4.

Method

Participants

Forty-eight members of the Society for Judgment and Decision Making agreed to take an online survey. Two participants were deleted since they did not answer all of the questions, bringing the sample size to 46.

Procedure

An email was sent out on the SJDM listserv asking for participants. A link was provided in the email to an online survey. Those who agreed to participate clicked on the link and provided consent. They were then presented with a brief online survey. The SJDM members were then presented with the same four replication scenarios that were used in Experiments 2 and 3. After each replication scenario, SJDM members were asked to estimate its transparency.

Materials

All data were collected via the software program MRinterview. After following the provided link, SJDM members were shown the following instructions:

Hello and thank you for participating. In this questionnaire you will be presented with four brief scenarios from SJDM literature. You will be shown the questions for each scenario as they would appear to research participants in a within-subjects design, in which the relevant comparisons to be made are made by the same research participants 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 participants. 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 participants. Here we will assume an experiment is “transparent” if it would seem to be relatively easy for your typical participant 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 participant 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.

The instructions and descriptions of scenarios that followed were identical to those in Experiment 3, save that SJDM members were not asked for any confidence ratings. All instructions and questions used in Experiment 4 appear in Appendix C.

Results

In addition to being asked to assess each replication’s transparency, the SJDM members also answered all questions for each scenario. It may be interesting, therefore, to compare the SJDM members’ responses for the replication scenarios to the undergraduates’ responses.

Replication data, undergraduates vs. SJDM members

Below is a comparison of the replication data for the undergraduates and for the SJDM members. For the undergraduates there is 99.9% power to detect differences of 20 percentage points and 55% power to detect such differences with the SJDM members. What is important, again, is whether the overall qualitative pattern of data is the same as the study being replicated and not whether statistical significance is achieved.

The marbles lotteries data, undergraduates (N = 256) vs. SJDM members (N = 46)

	Option A	Option B	Option C	Option D
Undergraduates:	6% (16)	94% (240)	70% (178)	30% (77)
SJDM members:	26% (12)	74% (34)	50% (23)	50% (23)

Figure 39. The marbles lotteries data, undergraduates vs. SJDM members

With the undergraduates, the 94% who preferred *B* and the 70% who preferred *C* were both significantly different from 50%, $z = 11.09$, $p < .05$ and $z = 4.62$, $p < .05$, respectively. With the SJDM members, the 74% who chose *B* was significantly different from 50%, $z = 2.37$, $p < .05$. In the other condition, however, 50% chose *C* and 50% chose *D*. Interestingly, a smaller proportion of SJDM members chose dominating Option *B* over *A* than the undergraduates. More of them, however, recognized that Option *D* dominates *C*. Thirty-seven percent of the SJDM members (17 participants) switched preferences between conditions. Of this 37%, 18% switched from preferring the dominated Option *A* to the dominating Option *D*, while 82% switched from preferring the dominating Option *B* to the dominated Option *C*. Interestingly, both switching from *A* to *D* and switching from *B* to *C* constituted statistically significant preference reversals. The 95% confidence interval for *A* and *D* is .07 to .29 and for *B* and *C* is .71 to .93. With the undergraduates, 68% (175 participants) switched preferences. Of this 68%, 3% switched from Option *A* to *D* and 97% from *B* to *C*. For *A* and *D* the 95% confidence interval is 0 to .08 and for *B* and *C* is .92 to 1.02. Given that Tversky and Kahneman

(1986) had hypothesized that participants would chose Options *B* and *C*, it is interesting SJDM members also exhibited a significant preference reversal in switching from *A* to *D*.

The Princeton applicant data, undergraduates (N = 256) vs. SJDM members (N = 46)

Forty-three percent of the SJDM members (20 participants) accepted the applicant with an average of *B*. Of this 43%, 20% (4/20) opted to wait for noninstrumental information. The 95% confidence interval for this 20% is .08 to .32, indicating that a disjunction effect was present in a significant proportion of participants' decisions. With the undergraduates, 84% accepted the applicant with an average of *B*. Of these, 48% (104 participants) chose to wait for irrelevant information. The 95% confidence interval for this 48% is .34 to .63.

The child custody case data, undergraduates (N = 256) vs. SJDM members (N = 46)

	Award <i>A</i>	Award <i>B</i>	Deny <i>A</i>	Deny <i>B</i>
Undergraduates:	70% (180)	30% (76)	28% (72)	72% (184)
SJDM members:	50% (23)	50% (23)	43% (20)	57% (26)

Figure 40. The child custody case data, undergraduates vs. SJDM members

Again, we will not bother calculating $P_c + P_r$ as it is not a useful way of looking at the data. This is because $P_c + P_r$ can be > 100 for Parent *B* even when it is not the case that most participants awarded and denied custody to *B* more than *A*. The undergraduates overwhelmingly preferred Parent *A* to have custody of the child in both conditions, contrary to Shafir's (1993) hypothesis. Though this was not the case with the SJDM members, their data also did not support Shafir's (1993) hypothesis, as they did not both award and deny custody to Parent *B* more than *A*. Fifty percent of SJDM members awarded custody to *A* and 50% to *B*. In the second condition, the 57% who denied custody to *B* was not significantly different from 50%, $z = .67, p > .05$. Thus, in both conditions there was no preference between parents.

Only 4% of the undergraduates awarded and denied custody to the same parent. (In Experiments 2 and 3. This was without the leading instructions found in Experiment 1). The 95% confidence interval for this 4% is 0 to .10, indicating that it is not significantly different from zero. Twenty-four percent of the SJDM members awarded and denied custody to the same parent. The 95% confidence interval for this 24% is .12 to .36, indicating that a significant proportion of SJDM members awarded and denied custody to the same parent.

The Asian disease problem data, undergraduates (N = 256) vs. SJDM members (N = 46)

	Program A	Program B	Program C	Program D
Undergraduates:	63% (161)	37% (95)	35% (90)	65% (166)
SJDM members:	80% (37)	20% (9)	57% (26)	43% (20)

Figure 41. The Asian disease problem data, undergraduates vs. SJDM members

With the undergraduates, the 63% who preferred Program A and the 65% who preferred D were both significantly different from 50%, $z = 2.97, p < .05$ and $z = 3.43, p < .05$, respectively. With the SJDM members, 80% chose Program A and 57% chose its equivalent Program C. The 80% was significantly different from 50%, $z = 3.02, p < .05$. The 57% was not, $z = .67, p > .05$. To look only at this data would, however, be misleading. Thirty-nine percent of the undergraduates (99/256) switched preferences between the survival and mortality frames. Of this 39%, 86% chose Programs A and D. The 95% confidence interval for that 86% was .76 to .96. With the SJDM members, 33% (15/46) switched preferences between frames. Of these, 87% chose Programs A and D. The 95% confidence interval for this 87% is .77 to .96. Thus, there was still a statistically significant preference reversal present in their data. Below is a comparison of the transparency ratings provided by the SJDM members to those of the undergraduates.

The SJDM members were asked to identify themselves as being either undergraduate students, graduate students, assistant professors, associate professors, full professors or

professors emeritus. A couple of SJDM participants contacted the author and said that they were post docs. They were told to identify themselves as assistant professors. Before we compare the SJDM members' transparency ratings to the undergraduate ratings obtained in Experiment 3, let's first take a look at whether there any such demographic differences in the SJDM sample.

Proportions of SJDM members who chose each transparency category by demographic group (N = 45)

Since one SJDM participant did not respond on this demographic question, the sample size for this section is 45, rather than 46. The percentages of SJDM members who fell into each category were as follows: 2% (1) undergraduate students, 47% (21) graduate students, 16% (7) assistant professors, 11% (5) associate professors, 22% (10) full professors and 2% (1) professors emeritus. Another way to look at this is that the sample was 49% (22) students and 51% (23) professors. Assuming that with 32% of participants correctly identifying the prediction that the Asian disease problem is "somewhat transparent," all demographic groups were accurate in their transparency ratings 50% of the time, with the single exception of the undergraduate student, who was 100% inaccurate. Since there was only one undergraduate student, however, nothing can really be concluded regarding undergraduate transparency ratings. The proportions of SJDM members who chose each transparency category by demographic group can be seen in the table below.

TABLE 6

SJDM MEMBER TRANSPARENCY RATINGS

		Completely	Somewhat	Not transparent
Undergraduates	Marbles lotteries	0% (0)	100% (1)	0% (0)
	Princeton app	100% (1)	0% (0)	0% (0)
	Child custody	100% (1)	0% (0)	0% (0)

	Asian disease	100% (1)	0% (0)	0% (0)
Graduates	Marbles lotteries	5% (1)	29% (6)	67% (14)
	Princeton app	19% (4)	52% (11)	29% (6)
	Child custody	57% (12)	29% (6)	14% (3)
	Asian disease	19% (4)	62% (13)	19% (4)
Assistant prof.	Marbles lotteries	0% (0)	43% (3)	57% (4)
	Princeton app	29% (2)	57% (4)	14% (1)
	Child custody	71% (5)	14% (1)	14% (1)
	Asian disease	14% (1)	86% (6)	0% (0)
Associate prof.	Marbles lotteries	20% (1)	20% (1)	60% (3)
	Princeton app	20% (1)	60% (3)	20% (1)
	Child custody	60% (3)	40% (2)	0% (0)
	Asian disease	0% (0)	100% (5)	0% (0)
Full professors	Marbles lotteries	0% (0)	40% (4)	60% (6)
	Princeton app	40% (4)	40% (4)	20% (2)
	Child custody	50% (5)	40% (4)	10% (1)
	Asian disease	20% (2)	50% (5)	30% (3)
Prof. emeritus	Marbles lotteries	0% (0)	100% (1)	0% (0)
	Princeton app	0% (0)	100% (1)	0% (0)
	Child custody	0% (0)	0% (0)	100% (1)
	Asian disease	0% (0)	100% (1)	0% (0)

Everyone was only 50% accurate, with the exception of a single undergraduate student, who was 0% accurate. Nothing, however, can be concluded regarding undergraduates based on a sample size of one. Thus, since no group was really any more or less accurate than any other, it is safe to collapse across these demographic groups.

Transparency ratings, undergraduates vs. SJDM members

Above, the N for the undergraduates was 256 because the data from both Experiments 2 and 3 could be used. For the transparency ratings, the N for the undergraduates will be 72 because here only Experiment 3 can be used for a comparison. The achieved power for the chi squares below is .84.

The marbles lotteries transparency ratings, undergraduates (N = 72) vs. SJDM members (N = 46)

	Completely	Somewhat	Not transparent
Undergraduates:	17% (12)	53% (38)	31% (22)
SJDM members:	4% (2)	37% (17)	59% (27)

Figure 42. The marbles lotteries transparency ratings, undergraduates vs. SJDM members

A chi square was run on the transparency ratings of the undergraduates and SJDM members and was significant, $\chi^2 (2, N = 118) = 10.45, p < .05$. 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.” Recall that despite this, only two undergraduate participants (3%) could correctly identify the research hypothesis for this scenario. Here, therefore, the SJDM members were accurate.

The Princeton applicant transparency ratings, undergraduates (N = 72) vs. SJDM members (N = 46)

	Completely	Somewhat	Not transparent
Undergraduates:	14% (10)	60% (43)	26% (19)
SJDM members:	28% (13)	50% (23)	22% (10)

Figure 43. The Princeton applicant transparency ratings, undergraduates vs. SJDM members

Here, the differences were statistically significant, $\chi^2(2, N = 118) = 3.70, p > .05$. A majority of SJDM members (78%) thought that the Princeton applicant within-subjects scenario would be at least “somewhat transparent.” Seventy-four percent of the undergraduates thought the same thing. Only one undergraduate, however, could correctly identify the research hypothesis for this scenario. Thus, neither the undergraduates nor the SJDM members were accurate here.

The child custody case transparency ratings, undergraduates (N = 72) vs. SJDM members (N = 46)

	Completely	Somewhat	Not transparent
Undergraduates:	25% (18)	51% (37)	24% (16)
SJDM members:	54% (25)	33% (15)	13% (6)

Figure 44. The child custody case transparency ratings, undergraduates vs. SJDM members

The undergraduate and SJDM members' transparency ratings were significantly different, $\chi^2(2, N = 118) = 10.11, p < .05$. This was because 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." In Experiment 3, it was similarly hypothesized that this replication would be the most transparent of all. In actuality, only five participants out of 72 could correctly identify the research hypothesis. Therefore, the undergraduates and the SJDM members (and the author) were inaccurate.

The Asian disease problem transparency ratings, undergraduates (N = 72) vs. SJDM members (N = 46)

	Completely	Somewhat	Not transparent
Undergraduates:	25% (18)	60% (43)	15% (11)
SJDM members:	17% (8)	67% (31)	15% (7)

Figure 45. The Asian disease problem transparency ratings, undergraduates vs. SJDM members

Here, the undergraduate and SJDM members' ratings were almost identical. All differences were nonsignificant, $\chi^2(2, N = 118) = 1.00, p > .05$. In Experiment 3 it was proposed that in order for a transparency rating of "somewhat transparent" to be accurate, roughly half of the sample should correctly identify the prediction in question. This was operationalized as the proportion of participants who could correctly identify the prediction not being significantly different from .5. Here, only 32% did, which was significantly different from 50%, $z = .84, p > .05$. According to this operational definition, both groups were inaccurate. Since there is no objective reason why "somewhat accurate" must be defined this way, and since many would likely argue that 32% of the sample correctly identifying the prediction should be considered "somewhat transparent," this interpretation will be adopted. Thus, here, both the undergraduates and the SJDM members were accurate. This results in the undergraduates being accurate 25% of the time and the SJDM members being accurate 50% of the time.

Discussion

An unspecified percentage comparison is to compare two or more percentages in a way that assumes either some degree of overlap or nonoverlap between them, sans any information regarding their actual degree of correlation (Lambdin & Burdsal, 2007). Because of this fallacy, reporting the percentage of participants who chose each option in an experiment is often insufficient. For example, in the Asian disease problem the data for the SJDM members was as follows: 80% chose Program A, 20% B, 57% C and 43% chose Program D. Looking only at a table of these percentages one would perhaps conclude that no preference reversal was present. This is because a majority of participants chose Programs A and C while a minority chose B and D. Nevertheless, 33% of the SJDM members switched preferences between the survival and mortality frames. This proportion is statistically significantly different from zero (the 95% confidence interval is .19 to .46). Thus, properly analyzing this data one would conclude that there was a significant preference reversal present.

The author believes this observation illustrates an important point regarding the reporting of results in JDM experiments. Because of the fallacy of unspecified percentage comparisons, to merely report averages or percentages is very often misleading; it can either hide interesting findings or make it look like something was found when really there is nothing there. This is an important point because presenting participants with options to choose from in multiple conditions and then presenting the results as the percentage of participants who chose each option is a standard practice in JDM literature. If no other information is given, such a reported result might qualify as what economist Thomas Sowell calls an “aha statistic” (Sowell, 1995). An “aha statistic” is when a selective representation of data paints one picture, when the inclusion of other information *from the same dataset* would paint a different picture. This is true whether displaying the proportions of participants who preferred each option in the Asian disease

problem, the proportions of participants who accepted the second gamble in a two-step gamble test of the disjunction effect or whether displaying the proportion of participants who chose and rejected each parent in the child custody case experiment. In all of these instances more data is necessary to paint a full picture.

Looking at the data from Experiment 4, in both the Asian disease problem and the marbles lotteries experiment, for both the undergraduates and the SJDM members there was a statistically significant preference reversal. In the Princeton applicant replication, the proportion of participants' decision in which a disjunction effect was present was statistically significant for both the undergraduates and the SJDM members. In the child custody case replication, only 4% of undergraduates awarded and denied custody to the same parent (95% confidence interval: 0 to .10). A significantly greater proportion of SJDM members (24%) awarded and denied custody to the same parent (95% confidence interval: .12 to .36). As has previously been the case, however, with the SJDM members the child custody case again failed to replicate. Thus, the same replication results were obtained for both the undergraduates and the SJDM members.

It is interesting that the same biases that affected the undergraduates' decisions also affected the decisions of the SJDM members, especially given that the entire motivation of using them as participants in Experiment 4 was the assumption that they were already familiar with these experiments. Given their results, however, they may not have been that familiar with these experiments after all. In almost every case, the effect obtained was significantly lower than the undergraduates' but still statistically significant. It is difficult, however, to draw any conclusions about the SJDM members, since we cannot say with any accuracy what they were thinking when they answered the replication questions. This then is a serious limitation with this experiment. How many of them were simply answering the questions? How many of them were answering

the questions the way they thought typical research participants would respond? It is difficult to say. If this experiment was re-run, it would certainly be helpful if the instructions were much more explicit regarding what the SJDM members should be doing when responding to the replication questions. Furthermore, it was not thought necessary to have the SJDM members try to identify the prediction for each replication, nor was it thought necessary to use a manipulation check. In hindsight, both would have been a good idea.

The main finding from Experiment 4 is that the SJDM members were unable to guess how transparent each within-subjects replication would be. Assuming that the Asian disease problem proved to be “somewhat transparent” (with 32% of the sample in Experiment 3 correctly identifying the prediction or manipulation), the SJDM members were accurate with the marbles lotteries and Asian disease problem, and were inaccurate with the other two replications. A 50% accuracy rate is akin to a coin toss; it is not demonstrative of any ability to correctly identify what will or will not prove to be transparent to research participants.

Again, it was hypothesized that SJDM members would tend to be familiar with the experiments being replicated and that many are also familiar with research methodology in general. It was further hypothesized that, because of this knowledge, the hindsight bias would influence their guesses, leading them to overestimate the transparency of each within-subjects replication. The data does not support this hypothesis. The SJDM members did not consistently overestimate the transparency of the replications. In fact, it was only with the child custody case replication that most SJDM members guessed that undergraduates would find it to be “completely transparent.” With the marbles lotteries, 59% guessed that undergraduates would rate it “not transparent.” Therefore, though it cannot be concluded that the hindsight bias impacted the SJDM members’ guesses, the data does illustrate that the SJDM members were

unable to accurately guess how transparent each replication would be to typical psychology research participants.

In conclusion, the SJDM members provided predictions about what would and would not be transparent to the typical psychology research participant. At least 50% of these predictions were inaccurate (debatably 75%); therefore, these predictions were not met. Since this was a sample of people who conduct research in the field of experimental psychology (as well as other fields), this provides evidence illustrating that researchers do not actually have any special insight into what will or will not prove to be transparent to research participants. This further suggests that the claim that in JDM experiments within-subjects designs should be avoided because they are transparent is likely unfounded.

CHAPTER 6

GENERAL DISCUSSION

Research summary

The results of Experiment 4 suggest that, contrary to what many authors have assumed, researchers actually have no special insight into what tasks will and will not prove to be transparent to research participants. Experiments 1 through 3 together suggest that the level of transparency actually achieved in some within-subjects studies of heuristics and judgments is not of a magnitude that results in confounding. Shafir's (1993) child custody case experiment consistently failed to replicate, but this failure was not due to within-subjects transparency. In Experiment 3, only 7% of participants could identify the research hypothesis and the same replication results were also obtained between subjects. Bastardi and Shafir's (1998) claim that replicating their effect within subjects tests whether it will hold when the task is rendered transparent was not supported. In Experiment 3, in the within-subjects replication of their Princeton applicant experiment, only 1% of participants were able to correctly identify the prediction when explicitly asked to do so.

Though transparency did not here result in demand characteristics, evidence was provided that the wording of instructions can indeed create very strong ones. In Experiment 1, when the instructions asked participants to "Please answer with whatever your first impulse is," 25% awarded and denied custody to the same parent in the child custody case replication. In Experiments 2 and 3 combined, with such instructions removed, only 4% of participants did so.

The replications of Shafir (1993) seem to have failed for two reasons. First, in almost every case participants preferred the impoverished option to have custody. Second, participants seem to have recognized that their responses should be complementary. Though it could be said

that this second reason indicates that at least one aspect of the within-subjects replication was transparent, it cannot be argued that this is in fact a difference in transparency between a within- and a between-subjects design. This is because in a between-subjects design, it does not make sense to expect that choosing and rejecting would be complementary in the first place. What would be in a within-subjects design one participant awarding and denying custody to the same parent, would be in a between-subjects design two different participants choosing different parents to get custody of the child. There is no reason to expect that the responses of two different people who have no knowledge of each other will be complementary.

One then should perhaps not predict that participants will do something as transparently irrational as awarding and denying custody to the same parent in a within-subjects design. Within-subjects transparency then may become a concern when one needs participants to respond in *saliently* illogical and irrational ways in order for the hypothesis to be supported. The word “salient” is stressed because it will likely be a case-by-case issue what does and does not prove to be saliently irrational. In short, it is an empirical question. In the within-subjects Princeton applicant replication, for instance, waiting for information that has zero bearing on the decision at hand did not prove to be saliently irrational. This does not mean, however, that we are “flying blind” when guessing what will prove to be salient. One would not, for instance, predict that participants would recognize that waiting for noninstrumental information in a disjunction problem would be an irrational thing to do—at least, not based on the evidence available. It has long been known that, for whatever reason, people perform surprisingly poorly when attempting to reason through disjunctive problems (e.g. Johnson-Laird & Wason, 1970; Sloutsky & Goldvarg, 2004).

An important implication of the present findings is that even when moderate transparency results from the within-subjects presentation of stimuli, this will likely still not have a deleterious effect on the data. Looking at the results of Experiments 1 through 3, Tversky and Kahneman's (1981) Asian disease problem replicated almost perfectly even though 32% of the sample was able to identify its research hypothesis in Experiment 3. This is perhaps evidence that a majority of participants would have to see through the research design in order for the results to suffer due to confounding. Furthermore, combining Experiments 2 and 3, the Asian disease problem replicated successfully despite the fact that only 39% of the sample (99/256), a minority, switched preferences between frames.

We can also look at the data from Experiments 2 and 3 to test the conjecture proposed in Experiment 1 that quite a few participants do not share the experimenter's interpretation of the stimuli. According to this conjecture, a number of participants interpret the statement, "400 people will certainly die," as meaning, "400 out of 600 will certainly die, maybe more." This is a perfectly valid interpretation of the description provided. Also according to this conjecture, a number of participants interpret the statement, "200 will be saved," as meaning, "at least 200 will live." As aptly pointed out by Kuhberger (1995), this statement does not tell one how many will in fact die. Out of 256 participants, 99 (39%) switched their preference between the survival and mortality frames. Of the 39% who switched, 86% (85/99) chose Programs *A* and *D* and only 14% chose *B* and *C*. A stark minority of those who switched chose *B* and *C*, which provides indirect support for the conjecture explained above. It does so because if one adopts these alternative interpretations of the stimuli, choosing Programs *A* and *D* becomes the most rational thing to do. It therefore must be mentioned again that to whatever extent participants held these alternative (and valid) interpretations of the stimuli, a framing effect cannot be said to have

occurred. If one adopts the alternative interpretations of Programs *A* and *D*, then choosing them is not a preference reversal but is rather the only way to maximize the number of innocent people saved. This then supports Kuhberger's (1995) point, who was perhaps the first to argue that the outcomes in Tversky and Kahneman's (1981) Asian disease problem are inadequately specified (see also Li & Xie, 2006).

In Experiment 2, the impact of order effects, differences in the data caused by displaying conditions together on the same page vs. separately on different pages as well as differences in the data caused by analyzing it both within and between subjects were tested for. Though small order effects may have been present, their impact was negligible. No order effects of a magnitude large enough to impact the overall pattern of results were found. One would likely expect that, because of transparency issues, analyzing the data both between and within subjects would result in greater differences in the data than would showing conditions on different rather than on the same page. This was not the case—presenting conditions on the same vs. different pages had a greater impact on the data than the alleged transparency of within-subjects designs. For two out of the three replications for which this comparison can be made (it cannot be made for the Princeton applicant replication), displaying conditions on the same vs. different pages resulted in greater differences than did analyzing it within vs. between subjects. Thus overall, a within-subjects analysis of the different-pages data resulted in slightly more dramatic results than those provided by a between-subjects analysis. With that said, however, in Experiment 3, presenting conditions together on the same page did not significantly increase task transparency. For the same-page data, out of 170 chances to guess the research hypothesis (34 participants x 5 scenarios—including the manipulation check), 25% of the guesses (42) were correct. For the

different-page data, out of 190 chances (38 participants x 5), 27% of the guesses were correct. This difference is not significant, $z = .43$, $p > .05$.

Implications for and discussion of common objections to the within-subjects design

Kahneman and Frederick (2005) argued that within-subjects designs should not be used for the study of heuristics and judgment, claiming that “factorial designs are transparent” (p. 280). It is certainly reasonable to assume that within-subjects designs must be—to some small degree at least—more transparent than their between-subjects counterparts. This does not, however, imply that the typical level of transparency achieved in within-subjects designs used to study heuristics and judgment is of a magnitude that would actually impact the results of the study.

The present results, for instance, suggest that the resulting transparency caused by presenting stimuli within subjects may not typically be severe enough to alter the overall pattern of data. Further, the results of Experiment 4 showed that three out of four experiments that SJDM members thought would be “somewhat” to “completely transparent” were in fact not, as demonstrated by the results of Experiment 3. Some researchers then may not have as much insight as they assume into how transparent many designs actually are.

It has also been argued that presenting JDM experiments within-subjects results in participants instantly recognizing the manipulation (Kahneman & Frederick, 2005). Supposedly, this creates an implicit message to participants that the researcher expects to find an effect for every manipulation recognized (Bar-Hillel & Fischhoff, 1981; Kahneman & Frederick, 2005). This argument is not supported by the present evidence. In Experiment 3, though participants were not explicitly asked to identify the manipulation of each experiment, every participant who did describe the manipulation was counted as having successfully completed the task. In each

case this was only a stark minority of the participants. In half of the cases the resulting proportion was not significantly different from zero. This admittedly grossly underestimates the proportion of participants who recognized the manipulation in each replication.

In the Asian disease problem, for instance, the manipulation should (one would think) almost pop out at the participant when presented within subjects. The crucial point, however, is that if we assume that most participants actually recognized the manipulation in all of the replications, it still does not appear that such recognition in any way confounded the results. This statement is supported by the fact that in Experiment 3, the exact same results were found *both within and between subjects for every replication*. In the child custody case replication, participants should presumably have an easy time recognizing the manipulation; but, as they by and large had no clue what was being predicted, no demand characteristics were created and the replication still failed. It could be argued then that recognizing the manipulation is not in fact enough to confound a study; some recognition of what is being predicted is also necessary. This would pose a quite serious confound. Though he was talking about predictions intentionally made public, the social sciences' most eloquent critic, Stanislaw Andreski, discussed the importance of hiding your predictions from those to whom they apply. If those the prediction applies to become aware of the prediction made, then self-fulfilling and self-negating effects creep into the equation, thereby reducing the social sciences to "the level of unrealistic pontification, no matter how refined its statistical techniques might be" (Andreski, 1972, p. 43).

Kahneman and Frederick (2005) also argued that, when presented within subjects, stimuli often become repetitive and impoverished and therefore participants begin responding in a mechanized fashion. Because of this, participants do not form an individualized impression of each stimulus. In response to this, first, if presenting a task within-subjects results in stimuli

becoming repetitive to the point that participants are mechanistically responding, then it could be argued that the real problem is the task is simply too long. The researcher may be trying to cram too much into one experiment. Second, though it is true that in a between-subjects design participants form “individual impressions” of the stimuli, such individual impressions, complete with their individual contexts, are often what result in between-subjects context effects (Birbaum, 1982, 1999; Parducci, 1965).

Another issue commonly taken with the within-subjects design is the argument that they lack ecological validity. This concern was discussed extensively in the Introduction. Briefly, if a between-subjects design is in fact the appropriate design to use given the hypothetical constructs involved, and if there is not a perceived threat of between-subjects context effects, then fine, use the between-subjects design. To irresponsibly argue, however, that between-subjects designs should exclusively be used in an entire field of research ignores the fact that between- and within-subjects designs are often not interchangeable. There are many instances in which the use of a between-subjects design is inappropriate if not completely invalid.

The above arguments against the use of within-subjects designs all have their merits. Furthermore, there are certainly additional objections to the use of within-subjects designs not covered here—indeed too many to include within the scope of this paper. They all, however, ignore the fact that between-subjects designs often produce bizarre, indeed, paradoxical results. Furthermore, when the results of a between-subjects design do not stand out as exceedingly weird, that does not necessarily mean they were not produced by the very same mechanisms that occasionally produce bizarre paradoxes. Therefore, they should be used with caution. Though the public loves hearing about exciting, counterintuitive and mind-bending findings as much as researchers love making them, whenever such exciting and counterintuitive results are obtained

between subjects our initial reaction should be one of skepticism. Within-subjects replication is necessary when possible.

Conclusion

The goal of this paper was to address two assumptions: first, that within-subjects designs are transparent and therefore should not be used to study heuristics and judgment; and, second, that between- and within-subjects designs are—issues of transparency aside—interchangeable. The first of these assumptions was addressed empirically. It was not found that transparency caused any changes in the results of any of the replications. More research is obviously needed, but if within-subjects designs do not in fact render most experimental tasks transparent, then within-subjects studies can be used more often without fear of confounding caused by demand characteristics. This would reduce cost and speed up research in the field. This is likely not a concern in most sub-disciplines within psychology or in other social sciences, where there is less resistance to the within-subjects design. In the field of JDM, however, there seems to be a bias toward using between-subjects designs because of the assumed transparency of their within-subjects counterparts. Because of this, it was important to test whether transparency actually does alter the results of within-subjects replications of between-subjects findings.

A limitation of this paper is the operational definition of “transparent.” Here, “transparent” was defined solely in terms of participants being able to discern what the experimenter was predicting. Another type of transparency that very well may pose a problem to some within-subject designs was discussed above. This is the transparency that may result when it becomes obvious to participants that they are being expected to act in irrational ways. As discussed above, when transparency is defined in this manner, one could conclude that the within-subjects replications of Shafir’s (1993) child custody case in Experiments 2 and 3 were in

fact “completely transparent,” since 96% of participants did not both award and deny custody to the same parent.

The second assumption addressed here was that within- and between-subjects designs are interchangeable. This was addressed by observing instances where between- and within-subjects designs do not test the same hypotheses. This could be the case because of the presence of between-subjects context effects or because of the particulars of the operational definitions employed. For whatever reason, whenever it is the case that between- and within-subjects designs do not test the same hypotheses, they cannot be said to be addressing the same questions. And when the two designs do not address the same questions they cannot be said to be interchangeable. It was argued that this is likely the case more often than commonly assumed.

Whenever in a between-subjects design the stimulus and context become confounded, interpretation of the results will at best be problematic and likely be impossible. Though I believe that more between-subjects studies suffer from such context effects than is typically appreciated, I do not go as far as Birnbaum (1999), who argued that between-subjects designs should never be used when the dependent variable is a subjective judgment. With that said, I recognize that much of this paper reads as though it is siding with Birnbaum. That is only because the issue of within-subjects transparency is both highly publicized and perhaps exaggerated, whereas Birnbaum’s argument regarding between-subjects contexts effects seems to be largely ignored and underappreciated. His argument, therefore, warrants some trumpeting while the other side of the aisle deserves a little scrutiny.

To summarize the points made in this paper, when choosing between within- and between-subjects designs there are two issues that should always be considered: context and design appropriateness given what exactly is being hypothesized. If a researcher is designing an

experiment in the areas of JDM, social psychology or any field in the social sciences, it is likely that the dependent variable will be subjective in nature. This does not necessarily imply that if a between-subjects design is used, that between-subjects context effects will confound the results. It does mean, however, that the possibility of their presence is immediately introduced. Thus, if the dependent variable is subjective and if a between-subjects design is being considered, the first issue to think about is what the perceived contexts in each group would be.

The researcher must remember that between-subjects designs limit the relevant information that participants are exposed to. This means that the context of the stimulus is intentionally being left unspecified. Participants cannot make decisions in a vacuum. To complete the task, they will try to fill in what the between-subjects presentation of stimuli is leaving out. They will, in short, automatically generate a context for the task at hand (e.g., Birnbaum, 1974, 1982, 1999). The experimental task cannot be completed otherwise. The question then is this: If you saw one condition, and one condition only, what context does that stimulus alone suggest? Is this context different than what the participants in the other condition are likely to conjure up?

The contexts could differ in two important ways. First, if the experiment involves participants ranking or rating stimuli of any kind, to complete such a task between subjects requires filling in the picture regarding the endpoints of the stimulus' scale. These endpoints will be subjectively generated in each condition. If the endpoints generated in each condition are different, then the resulting ratings or rankings are not comparable. Furthermore, specifying the scale is not enough to fix this problem. If judgments were only a function of the range principle, then the judgment functions themselves would be linear and specifying the scale in each group would solve the problem. Since judgments are also a function of the frequency principle,

however, their corresponding judgment functions can be nonlinear. Because of this, anchoring the range is not sufficient to produce comparable results (Birnbaum, 1974, 1999).

Second, contexts can differ between groups when the incomplete information imparted in between-subjects designs results in participants making different comparisons than would be made within subjects. In general, whenever experimenters wish to compare Likert-style ratings of stimuli, the comparisons that would be made between subjects might not be the same as the comparisons that would be made within; that is, unless participants in the between-subjects design are made aware of what the stimulus in the other condition is. Sometimes this will matter and sometimes it will not.

As noted in the Introduction, it is important to consider what participants might be thinking. If presenting stimuli within and between subjects results in different comparisons being made, which comparisons are relevant to the hypothesis? For example, student evaluations of teachers almost have to be conducted between subjects; but, doing so results in students comparing how well their teacher taught their present class to how well others taught other classes. This is really not all that pertinent to the assessment of a teacher's abilities. What would be relevant is how well that teacher did compared to how well other teachers could do teaching the very same curriculum (Birnbaum, 1999; Greenwald, 1997). Ignoring this point leaves teaching skill confounding with a multitude of things, and, as a result, student evaluations of teachers might not actually be all that informative or useful. If you wanted, however, to know how well participants liked using a Dell computer compared to an HP, it makes perfect sense to have Dell users rate their satisfaction with their computers and to do likewise with HP users. This is because it does not matter here if HP users end up comparing their HP model to other HPs rather than Dells—if, that is, there is a random sampling of models from both companies. If

the HP users tended to own a higher-end model whereas the Dell users tended to own a lower-end model, then confounds could be present.

In addition to context (both regarding differing scales and differing comparisons being made), one must also consider what exactly is being hypothesized. If what is being hypothesized is the presence of a bias strong enough to impact participants' decisions when comparing the relevant stimuli, then the only design that allows them to make these relevant comparisons is a within-subjects design. An example of this would be the disjunction effect (Tversky & Shafir, 1992). Similarly, testing for the conjunction fallacy requires a within-subjects design (Tversky & Kahneman, 1983). If, however, what is hypothesized is a bias that simply impacts responses regardless of whether or not participants are exposed to all relevant stimulus conditions, then a between-subjects test is appropriate. The wording of the operational definition of the effect in question goes a long way in determining whether one design is more appropriate than the other. Anchoring, for example, can be tested either between or within subjects (Chapman & Johnson, 1999). Testing how recall (Borgida & DeBono, 1989) or the impact of decisions aids or certain biases (Arnold, Collier, Leech & Sutton, 2004) differ between experts and novices requires a between-subjects design. Mental contamination studies (see Wilson & Brekke, 1994) often require a between-subjects design, as do studies where certain emotional states must be induced, such as some misattribution studies (e.g., Schachter & Singer, 1962).

Whether the reader agrees with the arguments made in this paper or not—and it is assumed that many will not given that this issue is, as Kahneman and Frederick said, “highly contentious” (2005)—it is hoped that this paper will nevertheless raise awareness of the issues discussed. The overall goal here is not ultimately to side with one camp or the other but to make it clear that choice of design type is a careful decision. When deciding if a between-subjects

design is appropriate one must consider what exactly is being hypothesized as well as what differences in context might exist between groups and how this might impact the interpretation of the results.

It seems to be a little appreciated point that *any* scientific finding is in part determined by the methods used to obtain it. This doubly so in the social sciences where dependent variables are often highly subjective and findings are highly dependent upon the selection of stimuli, how the hypothetical constructs involved are operationally defined, the experimental protocol employed, the design used, the statistical analyses chosen, the level at which the data are aggregated, etc. Because of the highly methods-dependent nature of findings, choice of design type should not be a casual decision. It is one of the most important steps in the execution of a study—and yet is one that is typically given little real thought. One cannot just assume that a between-subjects design is a viable solution to whatever perceived problems the use of a within-subjects design might pose. The evidence presented here suggests that one of these perceived problems—the assumed transparency of within-subjects designs—may actually be exaggerated, and also suggests that researchers actually have no special insight into what designs will or will not prove transparent to participants. There are pros and cons with both between- and within-subjects designs of course, though it seems that some of the between-subjects design's greatest cons are largely unrecognized by researchers.

REFERENCES

REFERENCES

- Andreski, S. (1972). *Social sciences as sorcery*. London, UK: Andre Deutsch Limited.
- Arnold, V., Collier, P. A., Leech, S. A. & Sutton, S. G. (2004). Impact of intelligent decision aids on expert and novice decision-makers' judgments. *Accounting & Finance*, Vol. 44(1), 1-26.
- Bagassi, M. & Macchi, L. (2006). Pragmatic approach to decision making under uncertainty: The case of the disjunction effect. *Thinking and Reasoning*, 12, 329-350.
- Bar-Hillel, M. & Fischhoff, B. (1981). When do base-rates affect predictions? *Journal of Personality and Social Psychology*, 41(4), 671-680.
- Bastardi, A. & Shafir, E. (1998). On the pursuit and misuse of useless information. *Journal of Personality and Social Psychology*, Vol. 75, No. 1, 19-32.
- Bazerman, M. H., Loewenstein, G. F. & White, S. B. (1992). Reversals of preference in allocation decisions: Judging an alternative versus choosing among alternatives. *Administrative Science Quarterly*, 37, 220-240.
- BirnbauM, M. H. (1974). Using contextual effects to derive psychophysical scales. *Perception & Psychophysics*, 15, 89-96.
- BirnbauM, M. H. (1982). Controversies in psychological measurement. In B. Wegner (Ed.), *Social Attitudes and Psychological Measurement* (pp. 401-485). Hillsdale, NJ: Erlbaum.
- BirnbauM, M. H. (1999). How to show that $9 > 221$: Collect judgments in a between-subjects design. *Psychological Methods*, 4(3), 243-249.
- 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.
- Borgida, E. & DeBono, K. G. (1989). Social hypothesis testing and the role of expertise. *Personality and Social Psychology Bulletin*, Vol. 15, no. 2, 212-221.
- Chapman, G. B. & Johnson, E. J. (1999). Anchoring, activation and the construction of values. *Organizational Behavior and Human Decision Processes*, 79(2), 115-153.
- Cialdini, R. B. (1984/1993). *Influence: The psychology of persuasion*. New York: William Morrow and Company, Inc.
- Cook, T. D. & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

- Dunning, D. & Parpal, M. (1989). Mental addition versus mental subtraction in counterfactual reasoning: On assessing the impact of personal actions and life events. *Journal of Personality and Social Psychology*, Vol. 57, No. 1, 5-15.
- 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*, Vol. 1, No. 3, 288-299.
- Fischhoff, B., Slovic, P. & Lichtenstein, S. (1979). Subjective sensitivity analysis. *Organizational Behavior and Human Performance*, 23, 339-359.
- Greenberg, G. & Lambdin, C. (2007). Psychology is a behavioral science, not a biological science: A discussion of the issue and a review of Neural Theories of Mind: Why the Mind-Brain Problem May Never Be Solved by William Uttal. *Psychological Record*, 57, 457-475.
- Greenwald, A. G. (1976). Within-subjects designs: To use or not to use? *Psychological Bulletin*, 83, 314-320.
- Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist*, 52(11), 1182-1186.
- Grice, G. R. (1966). Dependence of empirical laws upon the source of experimental variation. *Psychological Bulletin*, 66, 488-498.
- Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and Semantics: Vol. 3. Speech Acts* (pp. 41-58). New York: Academic Press.
- Hawkins, S. A. & Hastie, R. (1990). Hindsight: Biased judgments of past events after the outcomes are known. *Psychological Bulletin*, 107, 311-327.
- Hsee, C. K. (1996). The evaluability hypothesis: An explanation of preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes*, 46, 247-257.
- Johnson-Laird, P. N. & Wason, P. C. (1970). Insight into a logical relation. *Quarterly Journal of Experimental Psychology*, 22, 49-61.
- Jones, C. & Aronson, E. (1973). Attribution of fault to a rape victim as a function of respectability of the victim. *Journal of Personality and Social Psychology*, 26, 415-419.
- Kahneman, D. & Frederick, S. (2005). A model of heuristic judgment. *The Cambridge Handbook of Thinking and Reasoning*, New York: Cambridge University Press.
- Kahneman, D. & Tversky, A. (1979). Prospect theory: An analysis of decision under risk.

Econometrica, 47, 263-291.

- Kahneman, D. & Tversky, A. (1982). The simulation heuristic. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment Under Uncertainty: Heuristics and Biases* (pp. 201-208). New York: Cambridge University Press.
- Keren, G. (1993). Between- or within-subjects design: A methodological dilemma. In G. Keren and C. Lewis (Eds.), *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues* (pp. 257-272). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Kuhberger, A. (1995). The framing of decisions: A new look at old problems. *Organizational Behavior and Human Decision Processes*, 62, 230-240.
- Kuhberger, A., Komunska, D., & Perner, J. (2001). The disjunction effect: Does it exist for two step gambles? *Organizational Behavior and Human Decision Processes*, 85, 2, 250-264.
- Lambdin, C. & Burdsal, C. (2007). The disjunction effect reexamined: Relevant methodological issues and the fallacy of unspecified percentage comparisons. *Organizational Behavior and Human Decision Processes*, 103, 268-276.
- Li, S. & Xie, X. (2006). A new look at the “Asian disease” problem: A choice between the best possible outcomes or between the worst possible outcomes? *Thinking and Reasoning*, vol. 12, no. 2, 129-143.
- Loftus, G. R. (1991). On the tyranny of hypothesis testing in the social sciences. *Contemporary Psychology*, Vol. 36, No. 2, 102-105.
- Lykken, D. T. (1991). What’s wrong with psychology anyway? In D. Cicchetti & W. M. Grove (Eds.), *Thinking Clearly About Psychology; Volume 1: Matters of Public Interest* (pp. 3-39). Minneapolis: University of Minnesota Press.
- Mahoney, M. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research*, Vol. 1, No. 2, 161-175.
- Medvedev, Z. A. (1971). *The rise and fall of T. D. Lysenko*. Garden City, NY: Anchor Books.
- Mellers, B. A., Hertwig, R. & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science*, Vol. 12, No. 4, 269-275.
- Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review*, 72, 407-418.
- Parducci, A. & Perrett, L. F. (1971). Category rating scales: Effects of relative spacing and frequency stimulus values. *Journal of Experimental Psychology Monographs*, 89, 427-452.

- Savage, L. J. (1954). *The foundations of statistics*. USA: John Wiley & Sons, Inc.
- Schachter, S. & Singer, J. E. (1962). Cognitive, social and physiological determinants of emotional state. *Psychological Review*, 69(5), 379-399.
- Shafir, E. (1993). Choosing versus rejecting: Why some options are both better and worse than others. *Memory & Cognition*, 21, 546-556.
- Shafir, E., Simonson, I. & Tversky, A. (1993). Reason-based choice. *Cognition*, Vol. 49, 11-36.
- Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology*. USA: Basic Books, Inc.
- Sloutsky, V. M. & Goldvarg, Y. (2004). Mental representation of logical connectives. *Quarterly Journal of Experimental Psychology: Human-Experimental Psychology*, 57A(4), 636-665.
- Slovic, P. & Fischhoff, B. (1977). On the psychology of experimental surprises. *Journal of Experimental Psychology: Human Perceptions and Performance*, 3, 4, 544-551.
- Sorokin, P. (1956). *Fads & foibles in modern sociology*. USA: Henry Regnery Company.
- Sowell, T. (1995). *The vision of the anointed: Self-congratulation as a basis for social policy*. USA: BasicBooks.
- Sun, Y., Li, S. & Li Y. (2008). Reexamining the role of the description of texts in the disjunction effect. *The Journal of Psychology*, 142(3), 261-265.
- Taleb, N. (2007). *The black swan: The impact of the highly improbable*. New York: Random House.
- Thompson, V. A. & Campbell, J. I. D. (2004). A power struggle: Between- vs. within-subjects designs in deductive reasoning research. *Psychologia*, 47, 277-296.
- Tversky, A. & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, Vol. 211, No. 4481, 453-458.
- Tversky, A. & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, Volume 90, Number 4, 293-315.
- Tversky, A. & Kahneman, D. (1986). Rational choices and the framing of decisions. *Journal of Business*, vol. 59, no. 4, pt. 2, 251-278.
- Varey, C., Mellers, B. A. & Birnbaum, M. H. (1990). Judgments of proportions. *Journal of*

Experimental Psychology: Human Perception and Performance, 16, 613-625.

Zikmund-Fisher, B., Fagerlin, A. & Ubel, P. A. (2004). Is 28% good or bad? Evaluability and preference reversals in health care decisions. *Medical Decision Making*, 142-148.

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

APPENDICES

APPENDIX A

INSTRUCTIONS AND QUESTIONS USED IN EXPERIMENT 2

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:

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

Scenario 2:

Imagine you are on the admissions committee of Princeton University. You are reviewing the file of an applicant who plays varsity soccer, has supportive letters of recommendation, and is editor of the school newspaper. The applicant has a combined SAT score of 1250 and a high school average of B. Do you choose to:

____Accept the applicant?

____Reject the applicant?

Again, imagine that you are on the admissions committee of Princeton University. You are reviewing the file of an applicant who plays varsity soccer, has supportive letters of recommendation, and is editor of the school newspaper. The applicant has a combined SAT score of 1250 but you have two conflicting reports of the applicant's high school average grade. The guidance counselor's report indicates a B average, while the school office reported an A average. The school has notified you that the records are being checked, and that you will be informed within a few days which of the averages is the correct one. Do you choose to:

____Accept the applicant?

____Reject the applicant?

____Wait for clarification from the applicant's school before deciding?

If you chose to wait for clarification in the question above, please answer the following:

The school informs you that the applicant's average grade is a B. Do you choose to:

____Accept the applicant?

____Reject the applicant?

Scenario 3:

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 4:

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*

APPENDIX B

INSTRUCTIONS AND QUESTIONS USED IN EXPERIMENT 3

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:

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 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 you are on the admissions committee of Princeton University. You are reviewing the file of an applicant who plays varsity soccer, has supportive letters of recommendation, and is editor of the school newspaper. The applicant has a combined SAT score of 1250 and a high school average of B. Do you choose to:

____Accept the applicant?

____Reject the applicant?

Again, imagine that you are on the admissions committee of Princeton University. You are reviewing the file of an applicant who plays varsity soccer, has supportive letters of recommendation, and is editor of the school newspaper. The applicant has a combined SAT

score of 1250 but you have two conflicting reports of the applicant's high school average grade. The guidance counselor's report indicates a B average, while the school office reported an A average. The school has notified you that the records are being checked, and that you will be informed within a few days which of the averages is the correct one. Do you choose to:

____ Accept the applicant?

____ Reject the applicant?

____ Wait for clarification from the applicant's school before deciding?

If you chose to "wait for clarification" in the question above, please answer the following:

The school informs you that the applicant's average grade is a B. Do you choose to:

____ Accept the applicant?

____ Reject the applicant?

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:

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 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

Scenario 5:

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 5.

- 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 4

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 participants in a within-subjects design, in which the relevant comparisons to be made are made by the same research participants 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 participants. 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 participants. Here we will assume an experiment is “transparent” if it would seem to be relatively easy for your typical participant 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 participant 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:

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

Scenario 2:

Imagine you are on the admissions committee of Princeton University. You are reviewing the file of an applicant who plays varsity soccer, has supportive letters of recommendation, and is editor of the school newspaper. The applicant has a combined SAT score of 1250 and a high school average of B. Do you choose to:

____Accept the applicant?

____Reject the applicant?

Again, imagine that you are on the admissions committee of Princeton University. You are reviewing the file of an applicant who plays varsity soccer, has supportive letters of recommendation, and is editor of the school newspaper. The applicant has a combined SAT score of 1250 but you have two conflicting reports of the applicant's high school average grade. The guidance counselor's report indicates a B average, while the school office reported an A average. The school has notified you that the records are being checked, and that you will be informed within a few days which of the averages is the correct one. Do you choose to:

____Accept the applicant?

____Reject the applicant?

____Wait for clarification from the applicant's school before deciding?

If you chose to "wait for clarification" in the question above, please answer the following:

The school informs you that the applicant's average grade is a B. Do you choose to:

____Accept the applicant?

____Reject the applicant?

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 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 4:

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