Holbrook, A. L., Krosnick, J. A., & Pfent, A. M. (in press). Response rates in surveys by the news media and government contractor survey research firms. In J. Lepkowski, B. Harris-Kojetin, P. J. Lavrakas, C. Tucker, E. de Leeuw, M. Link, M. Brick, L. Japec, & R. Sangster (Eds.), Telephone survey methodology. New York: Wiley.

Potential Drawbacks of Attribute-Based Stated Choice Studies: Comments on Layton and Brown (1998) and Swait, Adamowicz, and Louviere (1998)

Jon A. Krosnick

Ohio State University

January, 1999

This paper was presented at the NOAA Workshop on the Application of Stated Preference Methods to Resource Compensation, Washington, DC. The author wishes to thank Catherine A. Heaney for her helpful comments on the manuscript. Correspondence should be addressed to Jon A. Krosnick, Department of Psychology, 1885 Neil Avenue, Columbus, Ohio 43210 (email: Krosnick@osu.edu).

Potential Drawbacks of Attribute-Based Stated Choice Studies: Comments on Layton and Brown (1998) and Swait, Adamowicz, and Louviere (1998)

In comparison to conventional survey research, recent NOAA-commissioned contingent valuation (CV) surveys conforming to the methods advocated by NOAA's blue-ribbon panel (Arrow, 1993) might seem remarkably inefficient. In a typical non-CV survey of public opinion, respondents are usually asked about 3 closed-ended questions per minute on average, allowing a researcher to ask a total of 135 questions on a wide range of topics in a 45-minute period. By contrast, CV surveys usually spend the same amount of time to ask respondents just one willingness-to-pay (WTP) question, accompanied by a small set of additional questions used to assess the quality of answers provided to the WTP item.

The additional time in CV surveys is occupied by lengthy presentations of information to respondents about a single instance of damage to the natural environment and a proposed plan to repair such damage or to prevent it from occurring again. The WTP item then can ask respondents whether they would vote for or against the plan at a specific dollar value if it were offered to voters on a referendum in an election. Different respondents can be randomly assigned to receive different dollar values, so a population's WTP can only be calculated once large groups of respondents were asked about different dollar values. Thus, it would appear, remarkably little information is collected from each individual respondent, making the technique seem quite inefficient.

It is therefore no surprise that various observers have considered alternative approaches worth considering. In particular, a number of possible approaches to increasing efficiency have been proposed in recent years. For example, instead of asking respondents a single closed-ended

referendum question with a single posted price, respondents can be asked an open-ended question inquiring about their maximum WTP (see, e.g., Mitchell & Carson, 1989). This reduces the number of respondents needed, because it avoids having to randomly assign different respondents to different dollar values. Second, instead of asking people to evaluate just a single set of damages and a single repair or prevention program, a survey can present many sets of damages and many programs, seeking reactions to them all (Adamowicz, Boxall, Williams, & Louviere, 1998). And perhaps even better yet, a survey can avoid asking for dollar values altogether and simply seek to identify compensation programs that would be acceptable and equivalent substitutes for particular sets of natural resource damages (Jones & Pease, 1997). This is appealing because it simplifies the process of translating the results of a CV survey into a policy plan for natural resource damage compensation.

The work of Swait, Adamowicz, and Louviere (1998) and Layton and Brown (1998a; 1998b) represent implementations of such alternative methods. Both investigations presented respondents with a series of different prevention or compensation plans and asked respondents to choose their most preferred. In addition, Layton and Brown (1998a) presented respondents with various different potential damage scenarios. For each scenario, higher prices were associated with plans that would prevent more damage, and respondents were asked to choose among them. Because the damage and prevention scenarios differed along just one dimension (i.e., number of feet the Front Range forests would retreat), Layton and Brown (1998a) could calculate WTP per foot of forest loss.

Swait et al. (1998) presented only a single damage scenario but offered respondents prevention and compensation plans that varied in myriad ways: in terms of tidal marsh acres damaged, numbers of animals killed, acres of purchased wetland and partially-developed areas

for restoration and preservation, acres of dune creation and stabilization, and acres of newly-built artificial reefs. Because Swait et al. (1998) also varied the alleged costs of the proposed plans, they, too, were able to assess WTP for specific benefits. If these more complex measurement techniques are effective at generating valid and reliable data, they constitute more efficient approaches to assessing natural resource values in damage assessment cases.

In this paper, I will raise some issues for thought about these new study designs. In particular, I will outline a series of reasons to wonder whether the apparent gains in information per dollar spent on data collection might come at a significant compromise in the quality of data obtained. And if this is so, it would presumably suggest that we reconsider whether the apparent advantages of stated choice methods truly represent advances. Put slightly differently, stated choice methods seem so appealing in principle as to be comparable to a "free lunch". But we all know there is no such thing as a free lunch. My purpose in these remarks is to suggest some ways to decide whether this appealing lunch is worth its price.

Although a number of different sorts of issues could and should be raised in addressing that question, my focus here is on a subset of them, involving sequencing and order effects. I begin below by calling attention to a relatively small finding in Layton and Brown's (1998a) results, a finding that seems a bit puzzling in light of the others they report. I then offer a suggestion of why this puzzling result might have been obtained, pointing out how stated choice methods may focus and direct subjects' attention and decision-making processes, causing the obtained results to be contingent on seemingly insignificant decisions made by the researchers in designing the procedures. That is, if the procedures were designed differently, in ways that should not matter in principle, the observed results might be quite different. After illustrating

this point by reviewing relevant psychological studies conducted during more than 100 years, I close with remarks about respondent burden.

Insensitivity to Scope?

My starting point in this discussion is a relatively small finding reported in a corner of Layton and Brown's (1998a) report: that respondents' willingness to pay for forest loss was insensitive to the time horizon over which the proposed forest loss would occur: 60 years versus 150 years. In contrast, WTP was sensitive to amount of forest loss, type of abatement program, and program cost. Why might this have been? Layton and Brown (1998a) consider and reject two possible alternatives: that they lacked sufficient statistical power to detect sensitivity to time horizon, and that respondents failed to notice the time horizon in the written materials.

Another possible explanation not mentioned in Layton and Brown's (1998a) report is that both 60 and 150 years may have seemed to be equivalently long times to respondents. In fact, in Layton and Brown's (1998a) focus groups, one or more individuals mentioned that both time periods seemed equivalent, in the sense that both extend beyond his/her/their life expectancies (Layton, 1998). If most or all respondents thought simply in self-focussed terms along these lines or simply viewed both time horizons as equivalently representing "a moderately long time", we would not expect WTP to vary accordingly. And this is certainly a possible reason for the null result.

However, there is yet another possible explanation as well, one that is particularly interesting to consider. A unique feature of time horizon in this study is that it was varied between respondents rather than within respondents. That is, each respondent was given only a single time horizon but was asked to consider various different amounts of forest loss, types of abatement programs, and program costs. Therefore, in order to make choices among the options

offered on each menu, respondents could effectively ignore the time horizon, which was the same for each option. Instead, respondents presumably focussed their attention on the factors that did vary across the options presented on each menu: amount of loss, program type, and program cost. If time horizon had varied across the options, respondents might well have paid attention to it and used it to make judgments.

If this account is correct, what should we make of it? One possibility is that failing to vary time horizon within respondents led to an underestimation of its impact on WTP (i.e., the conclusion that it has no impact). By varying the other factors, this account proposes, the study was accurately able to gauge their impact. However, there is another possibility as well: focusing respondent attention on an attribute by varying it within respondents may have led them to place disproportionately greater weight on it when making judgments than would be the case outside the experimental context.

Artifacts in Within-Subjects Experimental Designs

The lawyer-engineer problem. To illustrate this potential problem, consider a study conducted by Kahneman and Tversky (1973) on the use of base rates in making social judgments. In their study, respondents were presented with what has come to be known as the "lawyer/engineer problem". The instructions for it read as follows:

> "A panel of psychologists have interviewed and administered personality tests to 30 engineers and 70 lawyers, all successful in their respective fields. On the basis of this information, thumbnail descriptions of the 30 engineers and 70 lawyers have been written. You will find on your forms five descriptions, chosen at random from the 100 available descriptions. For each description, please indicate your probability that the person described is an engineer, on a scale from 0 to 100.

"The same task has been performed by a panel of experts, who were highly accurate in assigning probabilities to the various descriptions. You will be paid a bonus to the extent that your estimates come close to those of the expert panel."

Respondents were then given five of the following sort of descriptions to read and judge:

"Jack is a 45-year-old man. He is married and has four children. He is generally conservative, careful, and ambitious. He shows no interest in political and social issues and spends most of his free time on his many hobbies which include home carpentry, sailing, and mathematical puzzles."

Kahneman and Tversky (1973) found that people's judgments were very responsive to the content of these thumbnails – the more a target matched the stereotype of a lawyer, the more likely people thought it was that he was indeed a lawyer. This is not especially surprising.

More surprising was the fact that respondents seemed to ignore the base rate nearly completely. From the point of view of rationality, the base rate constituted useful information – if a target was selected at random from among 70 lawyers and 30 engineers, he was much more likely to be a lawyer than if the selection population contained only 30 lawyers and 70 engineers. But in fact, it didn't much matter whether respondents were told the population contained 30 or 70 lawyers – this had almost no impact on judgments, altering the average prediction by a mere 5%. From this, Kahneman and Tversky (1973) concluded that people "largely ignored" base rate information when making social judgments, instead focusing attention on more vivid and idiosyncratic information about each case they confronted.

In fact, however, this conclusion was wrong, because Kahneman and Tversky's (1973) results were artifactual rather than real. In their study, the base rate was varied <u>between subjects</u>, meaning that each respondent was randomly assigned to be told either that the population

contained 70 lawyers and 30 engineers or that the population contained 30 lawyers and 70 engineers. In contrast, the thumbnail sketches varied within subjects, meaning that each respondent was given multiple different sketches and was asked to judge each one. This made the thumbnail sketches more salient in respondents' perceptual fields, and therefore focussed their attention on the sketches, leading them to have greater impact on people's judgments.

Replications of Kahneman and Tversky's study have shown that the fewer thumbnail sketches a respondent is given, the more he or she relies on the base rate (Ajzen, 1977; Fischhoff, Slovic, & Lichtenstein, 1979; Gegerenzer, 1991; Schwarz, Strack, Hilton, & Naderer, 1991). And when respondents are given only one thumbnail sketch in a between-subjects design, they are fully responsive to the base rate (Gigeranzer, Hell, & Blank, 1988). Thus, it is not true that people generally ignore base rate information and instead focus on vivid information in making social judgments, as Kahneman and Tversky (1973) concluded. Rather, it appears, respondents focus their attention on types of information that the experimenter varies across trials and underweight or ignore information that is held constant across trials. This conclusion has been validated by similar studies comparing within-subjects to between-subjects designs with judgment problems other than the lawyer-engineer problem as well (e.g., Birnbaum, 1982, pp. 442-445; Birnbaum, & Mellers, 1983; Hinz & Davidson, 1993; Westbrook, 1991).

Conversational conventions. Why might this be? One possible answer comes from the work of linguists, which sheds light on the process of communication that unfolds between researchers and their respondents in questionnaire-based studies such as these. According to Paul Grice (1975, 1978), speakers in everyday conversation conform to a set of norms about what to say and how to say it, and listeners interpret utterances presuming that the speakers are conforming to those norms. The result is an efficient process of communication wherein

information can be expressed implicitly rather than explicitly, allowing more to be conveyed than merely the literal meanings of the words alone.

One of the “rules of communication” is the “figure/ground rule”, which states that when speakers communicate a package of information, they should provide background information first, to establish a context in which to express the foreground information that should be the focus of the listener’s attention (Clark & Clark, 1977, p. 79; Halliday, 1967). In Kahneman and Tversky's (1973) experiment, the base rate was presented initially as background information, and the case studies were presented as varying foreground information, with the variation between them directing respondent attention to that information (Schwarz et al., 1991). In the absence of such a direction (albeit unintentional) from the experimenter in a between-subjects design, respondents allocate weight to judgmental criteria differently.

Greenwald (1976) referred to such cases in which within-subjects designs create effects of variables that would not appear in between-subjects designs as manifestations of sensitization effects, whereby respondents’ attention is directed to variation across trials, thereby distorting obtained results. That is, respondents are alerted to what the experimenter is varying and make inferences about what they are expected to do with the varying information.

Enhanced perceptual accuity. Other studies have shown the same sort of difference in results between between-subjects and within-subjects studies of perception rather than judgment, and the cognitive mechanism at work is likely to have been different. For example, Zeskind and Huntington (1984) explored people’s ability to distinguish between low-risk and high-risk crying by infants. People were unable to do so when considering just one cry at a time (in a between-subjects design), as would be the case for parents listening to their infant cry on a single occasion in their homes. But when different crying patterns were presented to respondents in quick

succession (in a within-subjects design), people were indeed able to distinguish between the two types. Here, conversational conventions were probably not at work, but instead, people were better-able to compare and detect differences between stimuli that would have seemed equivalent if considered individually.

Range effects. Other evidence suggests that discrepancies between the results of within-subjects and between-subjects experiments can occur for other reasons as well: range effects, practice effects, and carry-over effects, all of which have led a number of authors to recommend being very cautious about within-subjects designs (Grice, 1966; Greenwald, 1976; Poulton, 1973, 1974, 1982). A range effect was demonstrated, for example, in a study by Kennedy and Landesman (1963) and reviewed by Poulton (1973). In that study, respondents were asked to perform physical manipulations on a table set at various different heights. Respondents were randomly assigned to two different groups, receiving a set of either high or low table heights, with two heights in common (labeled A and B). One group of respondents performed significantly better at height A than height B, and the other group performed significantly better at height B than height A. Thus, the conclusion one would reach about the relative advantage of these heights depends completely upon which other heights are offered for testing.

Poulton (1973) reviewed dozens of other such range effects in within-subjects designs. More generally, numerous studies have shown that the range of stimuli offered in a within-subjects experiment influence the ratings given to those stimuli (Birnbaum, 1974; Helson, 1964; Johnson & Mullally, 1969; Mellers & Birnbaum, 1982, 1983; Parducci, 1968, 1982; Parducci & Perrett, 1971; Restle & Greeno, 1970). Thus, the set of stimuli (perhaps arbitrarily) chosen to present determines the way respondents evaluate them. Consequently, results from such a study may not necessarily generalize to other judgment contexts.

Practice effects. Practice effects occur when respondents are asked to perform the same sort of judgment task repeatedly – the more they do it, the better they get at it (e.g., Smith, 1989). Therefore, the results of later trials may look different from the results of earlier trials not because of the experimental variations but rather simply because of practice effects changing how later judgments are made. The impact of practice effects on research conclusions can sometimes be eliminated by counterbalancing the order of tasks across respondents. That is, different respondents can be given tasks in different orders. However, as Greenwald (1976) explained, practice effects can sometimes interact with the effects of other variables, so counterbalancing does not always eliminate their effects.

An example of a practice effect appears in a study by Winman (1997), who explored the degree to which people fall prey to hindsight bias. This is the tendency to believe that one has always possessed knowledge that was in fact only recently acquired; by the same token, once a historical development has occurred, people believe it could easily have been anticipated in advance. Although this bias was clearly present in between-subjects experiments by Winman (1997) and many others (e.g., Fischhoff, 1975), the effect disappeared in Winman's (1997) within-subjects design, because repeated trials familiarized respondents with the task, improved the logical consistency of their performance, and thereby eliminated the reasoning fallacy. Thus, the apparent rationality of people's thinking can be exaggerated by the appropriate within-subjects design of sequenced tasks.

Interestingly, practice effects were of great interest when within-subjects designs were first used in psychological research, pioneered by Wilhelm Wundt (see Boring, 1953; Hothersall, 1984). Wundt's method for studying psychological processes involved bringing single respondents to the laboratory, training them to introspect and describe the sensations they

experienced when observing stimuli, and then exposing them to numerous stimuli to which to react. Wundt considered practice essential: some of his experimental procedures did not make use of the information a respondent generated until after he or she had first honed his or skills by performing 10,000 introspective observations. Although this method enjoyed widespread use in psychology for decades, it ultimately fell out of favor completely and disappeared, importantly because scholars concluded the results obtained were artifactual results of the extensive training respondents had received (Boring, 1953; Hothersall, 1984).

Carryover effects. Carryover effects occur when the nature of the first task presented to respondents induces them to adopt a particular judgmental strategy, which they then apply to all later tasks in the sequence they are given. Or, as Greenwald (1976) put it, carryover effects occur when the effect of a manipulation implemented early in a sequence has effects on later behaviors. For example, in a study by Levin, Johnson, and Davis (1987), respondents were asked to judge a series of gambles, presented one at a time. The first gamble presented was evaluated positively when framed in terms of gains and was evaluated negatively when phrased in terms of losses. But the framing of the initial gamble carried over to evaluations of subsequent gambles, even when these later gambles were presented with different frames. Thus, a within-subjects design can fail to document an effect observable in between-subjects designs because of carry-over from the seemingly-arbitrarily tasks presented to respondents initially.

Reasons to like within-subjects designs. Within-subject designs do have some supporters, who have made a variety of arguments in their favor (e.g., Plake & Wise, 1986). For example, many scholars recognize that within-subjects designs in principle afford more efficient statistical power to detect effects of manipulations (Baron, 1990; Sidman, 1960). This is because manipulation effects must appear over and above chance differences between experimental

groups in between-subjects design, whereas this latter source of error variance is not present in within-subjects designs. However, studies empirically testing this claim have sometimes found more statistically significant effects of manipulations in between-subjects designs (e.g., File, 1992), thus validating the notion that additional sources of (presumably systematic) error in within-subjects studies (due to context, carry-over, and other such effects) come along with the elimination of between-group chance variation in those designs.

Still others have argued that context effects occur in between-subjects designs just as they do in within-subjects designs, but in the former, the contexts are not known to researchers, because respondents bring their own contexts from prior life experiences (e.g., Birnbaum, 1982). Yet such extra-experimental contextual effects would impact any real-world decision task and are therefore appropriately influential on any experimental result. This means that such effects are present in within-subjects experiments as well and, because of random assignment to experimental condition, do not distort the results of between-subjects experiments. Finally, others have challenged the claim that within-subjects designs inherently suffer from range effects (Rothstein, 1974), but these particular defenses have been discredited (Poulton, 1974). In sum, then, when discrepancies appear between studies using between-subjects and within-subjects designs, there are many reasons to suspect artifacts may be at work in the latter.

Implications Regarding Attribute-Based Stated Choice Studies

All this sheds valuable light on Layton and Brown's failure to document an effect of time horizon. In their study, cost, feet of tree loss, and solution program were varied within subjects, and these manipulations had observable effects on respondents' choices. But time horizon was varied between respondents. Quite possibly, people were led to be sensitive to that which varied across objects being considered. The failure to utilize time horizon may in fact be real, in the

sense that people genuinely did not see a meaningful difference between 60 and 150 years. But it may also be that people were led to ignore time horizon by inappropriately focusing their attention on the attributes of choices that varied within a menu. That is, the menu format may lead people to place weight on the varying dimensions and ignore or underweight dimensions held constant or unstated but relevant.

What would happen if we augmented the design, varying time horizon within menus in addition to varying price, feet of tree loss, and solution program constant? That is, respondents could be told that global warming will most likely cause 600 feet of tree loss, but the time it takes for that damage to occur is not known with certainty and could vary from 60 years to 150 years. The same menu would be presented a number of times, varying in the time horizon. By calling attention to time horizon in this way, we might well find that people say they are willing to pay more for damage that would occur more quickly.

So are people sensitive to scope in the Layton and Brown (1998a) study? They appeared insensitive to the speed at which the damage would occur, suggesting that the length of time during which people would be deprived of the trees did not affect values. This seems odd and challenges the notion of sensitivity to this attribute of environmental damage. But if we varied time horizon within respondents and found willingness to pay to be greater for damage that occurs more quickly, would that mean people are indeed sensitive to this aspect of scope? The psychological literature reviewed above suggests that instead, the observed responsiveness of judgments could be an artifactual result of the within-subjects design instead of genuine sensitivity.

Interestingly, Layton and Brown (1998a) assert that their respondents are sensitive to the scope of the injuries, but the effects the authors use to justify this conclusion are not in fact what

most scholars mean by scope. The usual definition of scope sensitivity is willingness to pay more money for a bigger good. But here, Layton and Brown show that holding price constant, people choose a larger good rather than a smaller one. This is quite a different phenomenon, and a much easier test to pass. But because of the study's design, it is not possible to conduct any proper test other than the one involving time horizon. And that test failed to document an effect. So, in fact, I would suggest the study might best be viewed as offering no convincing evidence of sensitivity to scope as currently reported.

More generally, I would suggest that the within-subjects design employed by Layton and Brown (1998a) and by Swait et al. (1998) may well create or artifactually strengthen apparent effects of manipulated variables. The result will be the appearance of sensitivity to an attribute when in fact people may manifest no such sensitivity in isolated individual judgments made in the natural course of daily life. Next, I turn to another potential source of bias in the observed results: response order effects.

Response Order Effects

A great deal of research during this century documents that when questionnaire respondents are asked to choose among a list of alternatives, their responses can be affected by the order in which the choices are presented (for a review of this literature, see Krosnick & Fabrigar, in press). The results of such impact are called response order effects. When response alternatives are presented visually, as is the case in Layton and Brown's (1998) and Swait et al.'s (1998) studies, people are inclined toward selecting the first options presented. This is because people tend to evaluate options with a confirmatory bias, whereby they tend to generate reasons to pick an alternative rather than reasons not to pick it. After considering the first couple of options with a confirmatory bias, people tend to become cognitively fatigued as their short term

memories become cluttered with the reasons they have generated to select the initially-presented options. When in this situation, some respondents short-circuit their evaluation of the later options and simply select one of the initial ones (e.g., Krosnick, 1991; Krosnick & Alwin, 1987).

This bias is likely to have been present in responses to Layton and Brown's (1998) questionnaire and, because of other design decisions, is likely to have altered people's selections and, more importantly, their apparent sensitivity to particular dimensions. In each menu presented to respondents, price increased from left to right. If people had a bias toward the initially-presented options, this would enhance their apparent preference to pay less. Layton and Brown cited this preference to pay less as evidence of rationality, but it could in fact have amounted to artifactual suppression of WTP amounts.

Furthermore, feet of trees increased from left to right in each menu. If people were inclined to select initially presented options, then this would have suppressed the appearance of a preference for saving more trees. That is, the impact of tree feet on WTP would have been artifactually underestimated. Thus, the inevitable inequitable presentation of the options in a menu (because some must be listed before others) can lead either to over-estimation or under-estimation of the impact of a judgmental dimension.

It might appear that this problem can be solved by randomly assigning respondents to receive the menu choices in random orders that vary across respondents. That way, no response option is presented before any other option unfairly often when averaging across respondents. But that, of course, defeats a supposed advantage of within-subjects designs: to gain precise insights about each individual respondent's willingness to pay. Only when a researcher combines across respondents can WTP be assessed.

More importantly, order rotation does not solve the problem but instead introduces a new source of error variance. One might imagine that statistically controlling for presentation order and averaging across respondents will remove any bias due to order. But in fact, not all respondents are equally susceptible to response order effects (e.g., Krosnick, 1991). Susceptibility is multiply determined and is in general a function of respondent cognitive skills, respondent motivation to think carefully about questions, and the cognitive difficulty posed by the questions. Cognitive skill and motivation levels vary across respondents and are functions of such factors as genetics, personality, and training.

If we could document, measure, and control for all of these factors, we could eliminate the within-subjects impact of response order effects. But we cannot yet do so, so some unexplained error variance in menu selections will be the result of differential sensitivity to response order. This will suppress the apparent impact of systematically varied attributes (e.g., feet of trees) relative to what would be observed in a between-subjects design, where each respondent is presented with only one object to evaluate (so no order effects are present). Consequently, in menus such as those used in attribute-based stated choice studies, order effects seem to be an unavoidable source of bias in weight estimates. Employing between-subject designs (as in more conventional CV studies) avoids this particular problem.

Cognitive Burden

To close these remarks, I want to raise one final set of considerations, about cognitive burden. As various observers have pointed out, attribute-based stated choice methods such as those used by Layton and Brown (1998) and Swait et al. (1998) involve all the essential elements of a contingent valuation study, plus they add substantial burdens for respondents. I mentioned earlier that a traditional CV study presents one damage scenario to each respondent, one

proposed plan to fix or prevent it, and one posted price. Different respondents can be told of different damage scenarios to test for sensitivity to scope, and different respondents can be told different prices, so a population's mean or median WTP can be calculated.

Layton and Brown (1998) presented multiple damage scenarios, multiple repair plans, and multiple prices, but the burden on questionnaire designer and respondent are the same as in conventional CV studies. The damage must be explained, the solution plan must be explained, and people must decide what they prefer. Yet the volume of information being conveyed to each person must increase many fold if depth and clarity are to be maintained in a multi-object judgment task. If information volume is to be held constant, then an attribute-based stated choice study must necessarily describe the damage and/or solution plans in much less detail than a CV study devoted to only a single damage scenario and solution plan.

Consequently, the apparent value of the within-subjects approach comes at yet another cost: either greater information volume to be processed or more superficial information presentation. Either way, the cognitive burden for respondents is increased, because more information processing is more fatiguing, and more superficial information presentation requires that respondents fill in desired but unstated details in order to make judgments. All this burden falls upon a respondent when he or she is trying to understand the options presented on a menu, even before he or she tries to make a choice among them.

When respondents turn to the task of making choices, they confront even further cognitive burden in the state choice studies. Instead of simply comparing the non-financial costs and benefits of one solution plan against its financial cost, respondents must compare multiple options that vary in terms of multiple attributes. In complex decision tasks of this sort, an accumulating body of evidence from psychological research indicates that decision quality

improves substantially when the task is broken down for people into its elementary steps (Armstrong, Denniston, & Gordon, 1975; Krosnick & Berent, 1993; Neijens, 1987).

This process, called decomposition, walks respondents through the sequence of making each sub-decision individually. By simplifying the cognitive task to be performed at each step, people's final judgments appear to be more rational and systematic and less error-laden. The literature on decomposition suggests that the judgments people make in tasks gauging WTP are likely to be of higher quality when tasks require that only one option be evaluated and of lower quality when multiple options must be evaluated and compared to one another at once. This points us back to square one: the old-style CV study.

Conclusion

Stated succinctly, anything one considers a drawback of traditional contingent valuation studies with regard to the quality of data obtained is present in attribute-based stated choice studies as well. But these latter studies entail a unique set of added potential drawbacks that suggest data quality may be lower as the result of a variety of systematic biases. If the appeal of these latter studies is their cost-effectiveness, we should probably recognize that with the cost savings most likely comes decreased accuracy. Is the tradeoff worthwhile? That may be a decision best made by policy-makers, not researchers.

References

Adamowicz, W. L., Boxall, P., Williams, M., & Louviere, J. (1998). Stated preference approaches for measuring passive use values: Choice experiments and contingent valuation. American Journal of Agricultural Economics, forthcoming.

Ajzen, I. (1977). Intutitive theories of events and the effects of base-rate information on prediction. Journal of Personality and Social Psychology, 35, 303-314.

Armstrong, J., Denniston, W., & Gordon, M. (1975). The use of the decomposition principle in making judgments. Organizational Behavior and Human Performance, 14, 257-263.

Arrow, K. (1993). Report of the NOAA panel on contingent valuation. Federal Register, 58 (10), 4602-4614.

Baron, A. (1990). Experimental designs. The Behavior Analyst, 13, 167-171.

Birnbaum, M. H. (1974). Morality judgments: Tests of an averaging model. Journal of Experimental Psychology, 102, 543-561.

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

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.

Boring, E. G. (1953). A history of introspection. Psychological Bulletin, 50, 169-189.

Clark, H. H., & Clark, E. V. (1977). Psychology and language. New York: Harcourt Brace Jovanovich.

File, S. E. (1992). Effects of Lorazepam on psychomotor performance: A comparison of independent-groups and repeated-measures designs. Pharmacology Biochemistry and Behavior, 42, 761-764.

Fischhoff, B. (1975). Hindsight ≠ foresight: The effect of outcome knowledge on judgment under uncertainty. Journal of Experimental Psychology: Human Perception and Performance, 1, 288-299.

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

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "Heuristics and Biases". In W. Stroebe & M. Hewstone (Eds.) European Review of Social Psychology (Volume 2, pp. 83-115). New York: Wiley.

Gigerenzer, G., Hell, W., & Blank, H. (1988). Presentation and content: The use of base rates as a continuous variable. Journal of Experimental Psychology: Human Perception and Performance, 14 , 513-525.

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

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 3: Speech acts (pp. 41-58). New York: Academic Press.

Grice, H. P. (1978). Some further notes on logic and conversation. In P. Cole (Ed.), Syntaz and semantics 9: Pragmatics (pp. 113-128). New York: Academic Press.

Halliday, M. A. K. (1967). Notes on transitivity and theme in English. Part 2. Journal of Linguistics, 3, 199-244.

Helson, H. (1964). Adaptation-level theory. New York: Harper & Row. 1964.

Hinsz, V. G. & Davidson, D. J. (1993). Contextual influences of within-subjects designs on base rate type problems. Paper presented at the meeting of the Judgment and Decision Making Society, Washington, DC.

Hothersall, D. (1984). History of psychology. New York: Random House.

Johnson, D. M., & Mullally, C. R. (1969). Correlation-and-regression model for category judgments. Psychological Review, 76, 205-215.

Jones, C. A., & Pease, K. A. (1997). Restoration-based measures of compensation in natural resource liability statutes. Contemporary Economic Policy, 15, 111-122.

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

Kennedy, J. E., & Landesman, J. (1963). Series effects in motor performance studies. Journal of Applied Psychology, 47, 202-205.

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response order effects in survey measurement. Public Opinion Quarterly, 51, 201-219.

Krosnick, J. A., & Berent, M. K. (1993). Comparisons of party identification and policy preferences: The impact of survey question format. American Journal of Political Science, 37, 941-964.

Krosnick, J. A., & Fabrigar, L. R. (in press). Designing Questionnaires to Measure Attitudes. New York, NY: Oxford University Press.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. Applied Cognitive Psychology, 5, 213-236.

Layton, D. F. (1998). Personal communication. Washington, DC.

Layton, D. F., & Brown, G. (1998a). Heterogeneous preferences regarding global climate change. Paper presented at the NOAA Workshop on the Application of Stated Preference Methods to Resource Compensation, Washington, DC.

Layton, D. F., & Brown, G. (1998b). Application of stated preference methods to a public good: Issues for discussion. Paper presented at the NOAA Workshop on the Application of Stated Preference Methods to Resource Compensation, Washington, DC.

Levin, I. P., Johnson, R. D., & Davis, M. L. (1987). How information frame influences risky decisions: Between-subjects and within-subject comparisons. Journal of Economic Psychology, 8, 43-54.

Mellers, B. A., & Birnbaum, M. H. (1983). Contextual effects in social judgment. Journal of Experimental Social Psychology, 19, 157-171.

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

Mitchell, R. C., & Carson, R. T. (1989). Using surveys to value public goods: The contingent valuation method. Washington, DC: Resources for the Future.

Neijens, P. (1987). The choice questionnaire: Design and evaluation of an instrument for collecting informed opinions of a population. Amsterdam: Free University Press.

Parducci, A. (1968). The relativism of absolute judgment. Scientific American, 219, 84-90.

Parducci, A. (1982). Category ratings: Still more contextual effects. In B. Wegener (Ed.), Social attitudes and psychophysical measurement. Hillsdale, N.J.: Erlbaum.

Parducci, A., & Perrett, L. (1971). Category rating scales: Effects of relative spacing and frequency of stimulus values. Journal of Experimental Psychology, 89, 427-452.

Plake, B. S., & Wise, S. L. (1986). Dynamics of guessing behavior: Between-group versus within-group designs. Bulletin of the Psychonomic Society, 24, 251-253.

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

Poulton, E. C. (1974). Range effects are characteristic of a person serving in a within-subjects experimental design – A reply to Rothstein. Psychological Bulletin, 81, 201-202.

Poulton, E. C. (1982). Influential companions: Effects of one strategy on another in the within-subjects designs of cognitive psychology. Psychological Bulletin, 91, 673-690.

Restle, F., & Greeno, J. G. (1970). Introduction to mathematical psychology. Reading, MA: Addison-Westley.

Rothstein, L. D. (1974). Reply to Poulton. Psychological Bulletin, 81, 199-200.

Schwaz, N., Strack, F., Hiltin, D. & Naderer, G. (1991). Base rates, representativeness, and the logic of conversation: The contextual relevance of 'irrelevant' inforamtion. Social Cognition, 9, 67-83.

Sidman, M. (1960). Tactics of scientific research. Evaluating experimental data in psychology. New York: Basic Books.

Smith, E. R. (1989). Procedural efficiency: General and specific components and effects on social judgment. Journal of Experimental Social Psychology, 25, 500-523.

Swait, J., Adamowicz, W., & Louviere, J. (1998). Attribute-based stated choice methods for resource compensation: An application to oil spill damage assessment. Paper presented

at the NOAA Workshop on Application of Stated Preference Methods to Resource Compensation, Washington, DC.

Westbrook, R. D. (1991). The partial reinforcement effect in human classical conditioning. Dissertation Abstracts International, 51 (8-B), 4083.

Winman, A. (1997). The importance of item selection in "knew-it-all-along" studies of general knowledge. Scandinavian Journal of Psychology, 38, 63-72.

Zeskind, P. S. & Huntington, L. (1984). The effects of within-group and between-group methodologies in the study of perceptions of infant crying. Child Development, 55, 1658-1665.