Psicológica (2014), 35, 405-422.

Evaluating Cognitive Models at the Group Level

David J. Weiss*

California State University, Los Angeles, USA

Cognitive models are usually conceptualized at the individual level, but are often analyzed at the group level. The level at which analysis can be carried out is dictated by the experimental design, which traditionally has been chosen for pragmatic reasons. Defining the model at the group level allows the incorporation of individual difference variables, which are of interest to many researchers, into the model structure. The nested groups design, with participants nested under the classificatory variables, is appropriate when the model is defined at the group level. That design is illustrated with a study testing a multiplicative model of anticipated compliance, in which medical patients were grouped according to the symptoms and prognoses associated with their diseases.

Models of cognitive processes are commonplace within a subset of the behavioral research community. Almost always, the model describes individual cognition, but in most research traditions is tested using data averaged over respondents. A consequence of this practice is the difficulty of examining individual differences variables, which are often dear to the hearts of researchers. In this essay, I illustrate how the use of nested group designs can help to overcome this deficiency.

When a fully crossed factorial design with independent groups is employed, a key question is whether the observed pattern of cell means reflects a single cognitive process shared by the population or is an artifact of the averaging. That is, whose cognition does the model describe? The question entails discussion of the issue of level of analysis, an issue deeply connected to experimental design. In this paper, that discussion is carried out in the context of two research traditions that examine how people make

* David J. Weiss is a Professor Emeritus of Psychology at California State University, Los Angeles. I wish to thank James Shanteau and Warren Thorngate for valuable comments on a previous draft, and Sergio Masin for spotting a historical lapse on my part. Address correspondence to Dr. David J. Weiss, e-mail: dweiss@calstatela.edu

judgments, with cognitive models at their core. Particular emphasis is given to the functional measurement tradition, because researchers employing that methodology have given serious consideration to the issue.

The Heuristics and Biases Tradition

Within the heuristics and biases literature that ultimately led to one of psychology's proudest moments, the award of the 2002 Nobel Prize in Economic Sciences to Daniel Kahneman, there is a paradox. The theoretical models describe individual cognition, but the empirical tests are carried out at the group level. For example, in discussing risky choice, Kahneman and Tversky (1984) report percentages of participants who choose one option over another, but they theorize about an individual's subjective state. It is not only researchers working in the heuristics and biases tradition who shift the level of analysis in this way. Their predecessors (e.g., Phillips and Edwards, 1966) and critics (e.g., Gigerenzer & Goldstein, 1996) do so as well.

The reason for the paradox is undoubtedly that the critical parameter in a choice problem is a probability; how likely is the respondent to choose one object over another? Estimating probability of choice within individuals is challenging. Not only does the researcher require large numbers of trials in order to get stable estimates of response frequency, but also those trials ought to be independent. People have the annoying habit of remembering their previous responses, which wreaks havoc with the independence requirement. Accordingly, pragmatic researchers often elect to estimate individual probabilities via group proportions (Kahneman & Tversky, 1982). The implicit assumption is that the average mind is a composite of homogeneous individual minds[1]. There are two ways in which this assumption might be wrong. Individuals might employ different models, or individuals might employ the same model but with idiosyncratic parameters.

Although the problem is well appreciated at the level of theory (Narens & Luce, 1983), with infrequent exceptions (e. g., Edwards, 1955; Tversky, 1967), the assumption of homogeneity has largely gone untested. If the goal of the research is the practical one of predicting group behavior, then of course basing the prediction on the behavior exhibited by the sample

[1] Although I have not seen the composition spelled out, I suspect the researchers in this tradition do not envision participants developing a shared mental model of the sort that may arise when a team collaborates on a common task (Klimoski & Mohammed, 1995). Instead, I believe the view is that participants share a common endowment, what Titchener labeled the generalized normal adult human mind (Shanteau, 1999).

is sensible. But if the goal is to understand how people think about choices, then more individually focused examination may be in order.

The Functional Measurement Tradition

Functional measurement researchers also evaluate cognitive models, usually presented in algebraic form. The title chosen by Anderson (1978) for one of his summarizing papers, "Progress in Cognitive Algebra"[2], conveys the optimism as simple algebraic models were found to provide good quantitative accounts of a variety of cognitive processes. The simplest of these models, the additive model and the multiplicative model, were found to describe a host of judgmental phenomena in domains as diverse as psychophysics (Anderson, 1970) and social psychology (Anderson, 1971).

Among functional measurement researchers, data analysis has usually been carried out at the level of the individual subject. The model is viewed as localized within the individual. With factorial designs generating the stimulus combinations to be judged, analysis of variance has been the primary tool for model evaluation (Weiss, 2006). Each model makes predictions about the significance and nonsignificance of specific sources. Single-Subject designs have the potential to provide very powerful tests of the model, because differences between people do not contribute to the error term.

When data are analyzed individually, the models for the various respondents may be revealed to differ from one another. That poses the challenge of making sense of the discrepancies. If there are identifiable commonalities among subsets of respondents, post-hoc clustering (Bonds-Raacke, 2006; Hofmans & Mullet, 2013) can provide convenient description. However, without additional information about the respondents, there may be insufficient leverage to clarify the factors leading people to employ different models. Post-hoc analysis can be suggestive, but is often low-powered because chance outcomes are uncontrolled. In general, such analyses do not provide the reassurance afforded by statistically testing a prior hypothesis.

Within the functional measurement tradition, group analyses were occasionally carried out, usually in situations where it was not

---

[2] Anderson (1971, 1981) also refers to the approach as information integration methodology. I prefer the original term, which highlights the associated goals. Algebraic models of cognitive processes are also key elements in other research streams, such as conjoint measurement (Luce & Tukey, 1964) and multi-attribute utility measurement (von Winterfeldt & Edwards, 1986), that have generally placed less emphasis on statistical aspects of model testing.

experimentally feasible to present each individual with the multiple replications needed to provide sufficient analytic power for single subjects. Three kinds of group design have been used.

Most commonly, group analyses have employed fully-crossed repeated measures designs. Each of a relatively small number of subjects goes through one (or rarely, several) replication of the factorial design, with the analysis following standard guidelines. Interaction with subjects serves as the error term for each substantive source. For additive models, standard computer ANOVA programs can carry out model evaluation. Multiplicative models are more complex to analyze, because the contrast coefficients needed to compute deviations from bilinearity, that is, what remains of the interaction after extracting its linear x linear component, are based on subjective values (Anderson & Shanteau, 1970). In the repeated-measures design, error terms also need to incorporate these coefficients[3].

Kaplan (1971; Kaplan & Kemmerick, 1974) introduced a variant in which subgroups of participants responded to different subsets of the experimental design. This approach yields designs in which subjects are nested under experimental factors, a structural relationship to be discussed below.

The third kind of group design, the independent groups design, was generally avoided until its feasibility was demonstrated by Howe (1991). The independent groups design has some advantages by virtue of requiring only one response from each person. First, it takes much less time per participant. For respondents other than students, a brief commitment can render recruitment feasible. Second, there need be no concern that the respondent can envision the factorial design and attempt to impose consistency on the set of responses. Third, the experiment can include extremely vivid, highly memorable stimuli without regard for the risk of people remembering their previous responses. On the negative side, though, there is a danger that between-Subject variability will limit the statistical power needed to evaluate a model.

These group designs all conceptualize the cognitive model at the level of the individual subject. Everyone is presumed to employ the same model. The group analyses are an experimental necessity to allow model testing and parameter estimation. A group is assembled by the researcher as an approximation to a random subset of the population of interest. The model and the parameters extracted are seen as those of a typical member of

[3] There were some glitches along the way (Graesser & Anderson, 1974), but after a few years a correct analysis was derived and incorporated into a computer program, POLYLIN (Weiss & Shanteau, 1982).

the population from which the participants were drawn. This is the usual way psychologists regard data based upon averages (Weiss & Edwards, 2005).

Individual Differences

But of course, many researchers have a great deal of interest in individual characteristics. Classificatory variables such as gender, ethnicity, health status, and age form the bases for well-defined areas of inquiry. Investigators employing regression-based analytic approaches routinely present models featuring demographic variables. Although there are exceptions, generally cognitive modelers have ignored individual differences predictable from grouping characteristics. Perhaps one reason cognitive models have not become more popular is that mainstream psychologists do not see how the formalism can be applied to their primary concerns.

Nested Designs

The problem for the modeler is that the interesting grouping variables are not subject to experimental control; one cannot assign gender, ethnicity, and the like for the sake of the study. Rather the researcher must recruit people with specified characteristics, a process that is not only inconvenient but also loses the inferential rewards of random selection. A partial solution to the problem is to include manipulated variables in a design in which participants are *nested* under the grouping variables. The key property of these designs is that each participant appears in only one level of the substantive factor under which the "Subjects" factor is nested[4]. Accordingly, one cannot estimate the interaction between a nested source and the source under which it is nested. Figure 1 shows the relationship in its simplest form.

Using a nested design to explore the role of a grouping variable produces complexities in three domains: the theoretical, wherein the locus of the model must be clarified; the statistical, in which novel estimation procedures are required, and the experimental, because the grouping itself may be imprecise. These issues will be easier to envision in the context of

[4] Designs in which subjects constitute the nested factor are sometimes called "mixed" or "split-plot". As the computations for all nested designs follow the same rules, there is little reason to introduce a separate label (Weiss, 2006). An expository discussion of nested designs is presented in Chapter 11 of that text.

an application, so an experiment reported by Rundall and Weiss (1998) will be used as an exemplar.

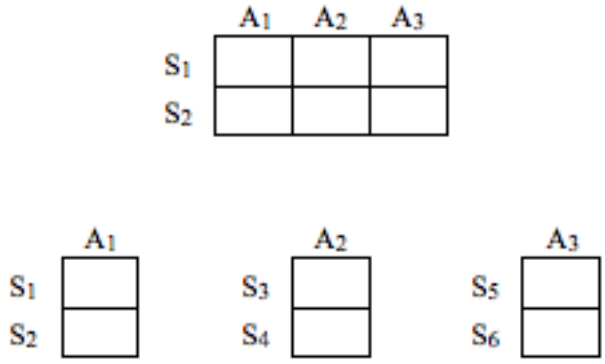

Figure 1. Top panel shows a crossed relationship between factors A and S (Subjects), in which all subjects responds to all levels of factor A. That relationship exemplifies the standard repeated measures design. The bottom panel shows a nesting relationship, in which Factor S is nested under Factor A. Note that the levels of the nested factor have subscripts that differ from column to column.

Rundall and Weiss (1998) asked 180 outpatients, each of whom had one of nine chronic diseases, to project their compliance in taking hypothetical prescribed medications specifically formulated for their diseases. Compliance is a critical problem in the treatment of long-lasting illnesses (Gerber & Nehemkis, 1986). Each of the five medications was described as likely to produce a particular side effect. The diseases used in the study were chosen to vary in symptom severity and disease prognosis, as specified by a medical manual, according to a 3 x 3 factorial design. With each medication rated twice, the design was a 3 (symptoms) x 3 (prognosis) x 5 (side effects) by 20 (patients) x 2 (replications). The "Patients" (i.e., subjects) factor was crossed with side effects and nested under symptoms and prognosis. That is, each patient projected compliance across all five of the side effects, but was limited (by actual medical condition) to only one combination of symptoms and prognosis. The nesting structure is shown in Table 1.

Table 1. Disease factorial design.

	PROGNOSIS		
	Favorable	**Doubtful**	**Unfavorable**
SYMPTOMS			
Asymptomatic or Mild	Iron Deficiency Anemia Patient 1 - Patient 20	Inactive Tuberculosis Patient 21 – Patient 40	Hypertension Patient 41 – Patient 60
Moderate	Hypothyroidism Patient 61 – Patient 80	Diabetes Mellitus(Type II) Patient 81 – Patient 100	Coronary Artery Disease Patient 101 – Patient 120
Severe	Epilepsy (Grand Mal Seizures) Patient 121 – Patient 140	Severe Bronchial Asthma Patient 141 – Patient 160	Chronic Obstructive Pulmonary Disease Patient 161 – Patient 180

The anticipation was that compliance would be greater for patients with more severe symptoms and with more dire prognoses, while troublesome side effects would interfere with pill-taking (Weiss, 1989) for everyone. A multiplicative model was hypothesized to describe quantitatively the manner in which the subjective impact of these variables combined to yield the projected compliance judgments:

$$R(i, j, k) = S(i) \times P(j) \times SE(k) + C_o \qquad (1)$$

In Equation 1, R(i, j, k) is the response. S(i) is the subjective value of the i-th level of symptoms. P(j) is the subjective value corresponding to the j-th level of prognosis. SE(k) is the subjective value of the k-th level of side effect. C_o is an additive constant providing an arbitrary zero point for the interval scale response.

The functional measurement school, like other cognitive modelers (e.g., Luce, 2010) regards a model as descriptive of the judgmental process, but does not claim to portray a participant's thoughts. People usually do not explicitly carry out the arithmetic suggested by the model equation when

making judgments or choices. When I compare offered pieces of cake with the intention of choosing the largest, I do not use geometry. My judgments are holistic, even though they might be well described by a Euclidean equation I learned in school. In experimental settings, the usual instructions request intuitive responses and attempt to inhibit calculation. Accordingly, we attribute to our formal models an "as-if" character. The quarterback aims the football as if he were using calculus to derive arcs and forces (Weiss, 2006; see also Pennings, 2003). The model is "paramorphic" (Hoffman's (1960) term conveying functional similarity) to the way in which the person integrates the subjective values of the components.

Locus of the Model

The model proposes that an individual's compliance is influenced by disease factors. That phrasing makes it sound like the model addresses an individual's cognitive process, but is misleading. Because no individual subject has all of the diseases, the model for compliance cannot be tested at the level of the individual, and therefore ought not to be defined at the level of the individual. The grouping variables themselves are model parameters. The model can be evaluated only at the overall group level. It is not a shared cognitive model, in the sense that medical anthropologists exploring cultural consensus examine beliefs shared by members of ethnic groups (Chavez, Hubbell, McMullin, Martinez, & Mishra, 1995). Rather, the model is descriptive of the cognitive behavior of the group as a whole, of a "group mind". In this case, the larger group to whom one might generalize consists of patients with the symptoms and prognoses specified in the factorial structure. This perspective allows specification of how aspects of personal history, encapsulated in this medical example as symptom severity and prognosis, affect the responses made by members of the sub-groups defined by the experimental design.

Statistical Issues

Because nesting eliminates estimation of interaction between a nested source and the source under which it is nested, the usual rule for repeated measures designs, wherein interaction with subjects generates the error term for each substantive source, leads to shared error terms. The composition of these error terms becomes increasingly complex as the number of nested sources increases, but packaged ANOVA programs have made the computations routine.

In testing an additive model, no additional complications arise with the use of nested designs. However, the test of a multiplicative model is a

different story. The difficult case arises when we explore a how a pair (or trio, etc.) of factors, at least one of which is nested under the Subjects factor, combines subjectively to influence responses. The derivation of error terms for polynomial components within a nested design, independent of the subjective spacing issue, is sufficiently complex that a well-known advanced text (Kirk, 1982) gives an incorrect formulation. A correct solution has been given by Myers (1979, p. 456). If the Subject factor is nested under Factor A (symptoms or prognosis in our example) and crossed with Factor B (side effects in our example), the mean square for S x lin(A)/B is the appropriate error term for all terms of the form p(A) x q(B), such as lin(A) x lin(B) or lin(A) x quad(B).

The FUNCTIONAL MEASUREMENT[5] computer program employs Myers's approach when a multiplicative model is specified for a nested design. Table 2 shows an excerpt from the Rundall and Weiss (1998) analysis of variance. The interaction term (shown in the first row) is decomposed to evaluate the hypothesis. The Side Effects and Prognosis factors were deemed to combine multiplicatively, because the bilinear (linear x linear) component of the interaction was significant and the other components were not. Note the use of the common error terms for the lin(SE) x lin(P) and lin(SE) x quad(P) terms, and for the quad(SE) x lin(P) and quad(SE) x quad(P) terms, etc. The sharing is called for when cycling through the levels of the nested source, Prognosis.

The solution proposed by Kirk (1982, p. 521) yields smaller error terms, and thus larger F-ratios, than that of Myers (1979). The Myers method is preferable because it provides an orthogonal decomposition. The sum of the sums of squares of the individual component error terms is equal to the sum of squares of the error term for the overall interaction. In a nonorthogonal decomposition, some of the variance in the data gets lost.

An alternative to the multilinear analysis has been used by other investigators. Tversky and Russo (1969) proposed evaluating multiplicativity by logarithmically transforming responses prior to carrying an additivity analysis. Inspired by the slide rule, this technique provides a plausible view of the pattern of the cell means, although the concomitant adjustment of the error variance may not be acceptable to some researchers. While the generality of the approach may be questioned, as responses with negative signs cannot be transformed, its use has persisted (Stevenson, Busemeyer, & Naylor, 1990), perhaps because the analysis may be carried

[5] FUNCTIONAL MEASUREMENT is included in the CALSTAT package of programs that accompanies Weiss's (2006) text.

out with standard ANOVA packages; no estimation of coefficients is required.

Table 2. Analysis of Variance for Side Effects (5 levels) x Prognosis (3 levels) interaction.

Source	*df*	SS	MS	F	p
(Side Effects x Prognosis	8	24333.60	3041.70	2.42	.014*)
Linear(SE) x Linear(P)	1	11428.18	11428.18	7.90	.006*
Linear(SE) x Quad(P)	1	2012.87	2012.87	1.39	.238
Error	171	247347.77	1446.48		
Quad(SE) x Linear(P)	1	2822.59	2822.59	2.02	.154
Quad(SE) x Quad(P)	1	2660.21	2660.21	1.90	.166
Error	171	239496.41	1400.56		
Cubic(SE) x Linear(P)	1	.10	.10	<1	.986
Cubic(SE) x Quad(P)	1	13.52	13.52	<1	.906
Error	171	179808.23	1051.51		
Quartic(SE) x Linear(P)	1	2999.29	2999.29	2.65	.101
Quartic(SE) x Quad(P)	1	2396.84	2396.84	2.12	.143
Error	171	193424.69	1131.14		

However, when the additive constant, C_o, differs across individuals, the transformational approach breaks down for group data. This can be illustrated by considering a group analysis of artificial data from two "Subjects" who are following the multiplicative model perfectly, but have different parameters. The "score" in each cell is the product of the individual's row "subjective value" and column "subjective value", plus the additive constant. First we compare the two analytic methods when $C_o = 0$.

The bilinearity test shows all of the interaction to be contained within the Linear x Linear component, as it should be, with F = 1.16. The column and row F-ratios are 1.59 and 1.20 respectively. The transformation

approach yields appropriate results as well. The logarithmically transformed responses yield an interaction F of 0, as they should for perfect data, with column and row F-ratios of 43.75 and 2.32.

Table 3. Artificial multiplicative data with $C_o = 0$

S1 $C_o = 0$

Subjective values	1	2	3
2	2	4	6
3	3	6	9
4	4	8	12

S2 $C_o = 0$

Subjective values	2	3	6
1	2	3	6
10	20	30	60
25	50	75	150

Thus, both methods report perfect multiplicativity for perfectly multiplicative data. However, a different result obtains when a varying additive constant enters the fray:

The bilinearity test yields exactly the same result as was obtained previously. All of the interaction is contained within the Linear x Linear component, with F = 1.16, while the column and row F-ratios are again 1.59 and 1.20. On the other hand, the test of the additivity of logged responses yields an interaction of 2.53, with column and row F-ratios of 46.41 and 2.32.

The conclusion is that when C_o varies, the bilinearity test still responds appropriately, in that it reports perfect multiplicativity when it should do so because all (in this illustration, both) subjects are multiplying, while the additivity test of transformed responses does not. With real subjects, the values of C_o are not predictable; the values may depend on idiosyncratic uses of the response instrument. It would seem safer to rely upon an evaluative method that is not disrupted by such differences.

Table 4. Artificial multiplicative data with varying C_o values

S1 $C_o = 2$

Subjective values	1	2	3
2	4	6	8
3	5	8	11
4	6	10	14

S2 $C_o = 5$

Subjective values	2	3	6
1	7	8	11
10	25	35	65
25	55	80	155

Experimental Issues

Grouping people according to natural labels is inevitably hazardous. The researcher is usually forced to accept group membership as reported; if reports are inaccurate, the group assignment is incorrect. Even more fundamental, perhaps, is the problem of category breadth. Do two people with a particular disease experience the same symptoms? Inevitably people whose subjective values may be quite disparate are going to be placed within the same group by an experimenter who is not omniscient. The statistical power to test the model is reduced by this variability.

Confounding is another potential issue with naturally occurring groups. Consider two of the asymptomatic diseases used by Rundall and Weiss (1998), inactive tuberculosis and hypertension. According to the factorial structure, these diseases vary in prognosis, and so differences between compliance estimates for the pair are attributed to the impact of that element. While the disease characteristics are accounted for by the experimental design, the patients who have the diseases differ in some important ways that the design does not capture. Inactive tuberculosis is a disease of the young; most patients are under 25. Hypertension, on the other hand, seldom appears before middle age. Older patients are especially predisposed to problems with medication compliance (Richardson, 1986). Tuberculosis patients are almost exclusively economically disadvantaged immigrants, while hypertension does not discriminate on the basis of wealth - but it is more likely to strike African Americans than member of other

racial groups. Might social variables also play a role in compliance (Castro et al., 1986)? Similarly, iron deficiency anemia is primarily a disease of women, while chronic obstructive pulmonary disease is largely a male concern. Might gender play a role in compliance (Connelly, Davenport, & Nurnberger, 1982)? Particular side effects, such as weight gain, might be seen as more burdensome by one gender rather than another.

The researcher may attempt to control known or suspected confounding by including additional factors in the design, but the increased complexity threatens to make the study infeasible. Rundall and Weiss (1998) did not have an easy time finding diseases that could be partitioned factorially according to symptoms and prognosis; additional constraints would have made a factorial structure impossible to achieve.

With the specialized requirements of a factorial design, recruiting appropriate participants can be a challenge. Additional factors increasingly restrict the number of people who qualify for a particular subgroup. Inevitably, some cells are easier to fill than others. For example, because hypertension is very common, Rundall and Weiss (1998) had no trouble finding patients with that disease. Inactive tuberculosis, on the other hand, is relatively rare, and access to patients through the normal channel of physician referrals proved fruitless. An additional issue is that some groups, such as patients with sexually transmitted diseases, may feel stigmatized, and the potential recruit may not be comfortable participating in the research specifically as a member of that group.

There are two ways to treat the problem of unequal availability. Rundall and Weiss (1998) produced equal cell sizes by recruiting as many subjects in all cells as the number attainable for the group that was most difficult to recruit. Equality of cell sizes kept the statistical analysis relatively straightforward. The alternative approach is to allow group sizes within the study to mirror proportions within the population. Proportional designs are appealing because they can allow increased generalizability when small subgroups that might otherwise be ignored are included in the research. Proportional designs yield orthogonal decomposition of variance. For an additive model, the consequent statistical complexity is not too great; for a multiplicative model, the analytic procedures for proportional designs have not yet been worked out.

Comparison of Design Types

Single-Subject analysis has been the primary vehicle for determining whether an algebraic model describes a set of judgments. Participants can be well trained, and the researcher can thereby feel confident that the

judgments are informed. Single-Subject data is generally characterized by low variability, which provides the analytic power needed to evaluate the model. Additionally, since there is no necessary reason for different individuals to employ the same model, the individual approach is seen as the appropriate way to describe behavior.

If inter-individual consistency happens to occur, all the better as it allows a simple description across people (Weiss & Anderson, 1969); but such consistency is not crucial to the success of the inquiry with regard to clarifying the cognitive process. The risk in merely assuming that everyone follows the same model was underscored by Shanteau and Anderson (1969), who found that an additive model of preference judgments fit group data quite well, but did not fit the individual data for 5 of 20 participants. Furthermore, a group analysis of just the minority who showed interactions did not yield a significant interaction. This tells us that the deviations from the model were idiosyncratic.

On the other hand, single-Subject designs are subject to carry-over effects; presentation order affects judgments. As Grice (1966) demonstrated and Poulton (1973) reiterated, substantive inferences can be markedly different from otherwise identical experiments using within-Subject designs versus between-Subject designs. Careful procedure can minimize the impact of carry-over effects, but only an independent groups design eliminates the concern. Carry-over effects such as stimulus contrast and response anchoring are interesting in their own right, and are certainly worthy of systematic study (Asch, 1946; Birnbaum, Parducci, & Gifford, 1971; Kaplan, 1971), but if uncontrolled can cloud effects of central interest. In the evaluation of a model such as that given by Equation 1, one that specifies the response to be a function only of the factorial combination, carry-over effects can be a serious intrusion.

A different kind of carry-over effect arises when subjects make many judgments. Complete factorial designs, especially with repetitions, can be tedious to evaluate. It is easy to neglect the participant's focus (Slovic, Lichtenstein, & Edwards, 1965). Boredom is a carry-over effect that has been generally overlooked by cognitive modelers.

The antidote to boredom presents carry-over risks of another kind. It would seem desirable to present stimuli capable of engaging the participant's interest. A consequent difficulty may be that vivid, memorable stimuli themselves call attention to the focus of the study. Participants exposed to more than one condition may guess the researcher's intention, and their personal theories may influence the responses in unaccountable ways.

For example, Harris & Weiss (1995) asked students to judge culpability in several ambiguous acquaintance rape situations. Pilot work showed it was not feasible to use scenarios in which the location of the protagonist couple's initial encounter (a bar or a coffee shop) was varied, because the students easily saw what was being manipulated. Although instructions called for global impressions, the respondents could not avoid formulating hypotheses about location. Demand characteristics were clearly coloring the responses. Unaware of the virtues of the nested design, the researchers elected to bypass the problem by fixing the location. Nesting respondents under location would have allowed exploration of an interesting variable.

Conclusion

Each kind of design a researcher might employ has advantages and disadvantages. The single-S design and the repeated-measures design are economical and efficient. With practice, an individual's use of the response scale is likely to stabilize. The negative feature is that repeated trials can lead to a variety of carry-over effects.

The independent groups design calls for recruiting a large number of participants, who are likely to vary in motivation and understanding, and whose approaches to the judgmental task may be quite idiosyncratic. Participants may use the response instrument idiosyncratically as well (Birnbaum, 1999). Response variability reduces the power of the statistical evaluation. When variability is large, it may not be possible to reject any model. An independent groups design requires an easy-to-use response instrument, since there will be no opportunity to learn.

Nested designs, in which each respondent sees only a specific subset of the stimulus combinations, afford the possibility of an intermediate position. In the functional measurement context, the nested design also deserves attention because of its potential value in exploring the kinds of demographic variables that are of wide interest. Nesting offers the researcher a principled way to investigate diversity among subjects. A nested group design might be applicable when personal characteristics, such as intelligence, experience, or wealth are embedded in a cognitive model. Or a nested design could be used to confirm characteristics suggested by a post-hoc cluster analysis of data collected using a single-S design.

The substantive conclusion of the Rundall and Weiss (1998) study is that anticipated compliance depends multiplicatively on disease aspects and side effects. That conclusion could not have been drawn from an analysis conducted at the level of the individual, nor could it have been drawn using

an independent groups design. Nested group designs can provide a means to extend the advantage of the carefully specified, quantitative, model to empirical questions whose exploration has previously been guided by verbal models (Harris, 1976).

REFERENCES

Anderson, N. H. (1970). Functional measurement and psychophysical judgment. *Psychological Review, 77*, 153-170.

Anderson, N. H. (1971). Integration theory and attitude change. *Psychological Review, 78*, 171-206.

Anderson, N. H. (1978). Progress in cognitive algebra. In L. Berkowitz (Ed.), *Cognitive Theories in Social Psychology* (pp. 1-126). New York: Academic Press.

Anderson, N. H. (1981). *Foundations of information integration theory*. New York: Academic Press.

Anderson, N. H., & Shanteau, J. C. (1970). Information integration in risky decision making. *Journal of Experimental Psychology, 84*, 441-451.

Asch, S. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology, 41*, 258-290.

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

Birnbaum, M. H., Parducci, A., & Gifford, R. K. (1971). Contextual effects in information integration. *Journal of Experimental Psychology, 88*, 158-170.

Bonds-Raacke, J. M. (2006). Using cluster analysis to examine husband-wife decision making. *The Psychological Record, 56*, 521-550.

Castro, F. G., Baray-Losk, A., McCreary, C., Cervantes, R., Bolden, D., Shieh, B., et al. (1986). Rehabilitation compliance in hand-injured Latino immigrant laborers: A multivariate stress-coping model analysis. *Journal of Compliance in Health Care, 1*, 111-133.

Chavez, L. R., Hubbell, F. A., McMullin, J. M., Martinez, R. G., & Mishra, S. I. (1995). Structure and meaning in models of breast and cervical cancer risk factors: A comparison of perceptions among Latinas, Anglo women, and physicians. *Medical Anthropology Quarterly, 9*, 40-74.

Connelly, C. E., Davenport, Y. B., & Nurnberger, J. I. (1982). Adherence to treatment regimen in a lithium carbonate clinic. *Archives of General Psychiatry, 39*, 585-588.

Edwards, W. (1955). The prediction of decisions among bets. *Journal of Experimental Psychology, 50*, 201-214.

Gerber, K. E., & Nehemkis, A. M. (1986). *Compliance: The dilemma of the chronically ill*. New York: Springer.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103*, 650-670.

Graesser, C. C., & Anderson, N. H. (1974). Cognitive algebra of the equation: Gift size = generosity x income. *Journal of Experimental Psychology, 103*, 692-699.

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

Harris, L. R., & Weiss, D. J. (1995). Judgments of consent in simulated rape cases. *Journal of Social Behavior and Personality, 10*, 79-90.

Harris, R. J. (1976). The uncertain connection between verbal theories and research hypotheses in social psychology. *Journal of Experimental Social Psychology, 12*, 210-219.

Hoffman, P. (1960). The paramorphic representation of clinical judgment. *Psychological Bulletin, 57*, 116-131.

Hofmans, J., & Mullet, E. (2013). Towards unveiling individual differences in different stages of information processing: a clustering-based approach. *Quality and Quantity, 47*, 455-464.

Howe, E. S. (1991). Integration of mitigation, intention, and outcome damage information by students and circuit court judges. *Journal of Applied Social Psychology, 21*, 875-895.

Kahneman, D., & Tversky, A. (1982). On the study of statistical intuitions. *Cognition, 11*, 123-141.

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist, 39*, 341-350.

Kaplan, M. F. (1971). Dispositional effects and the weight of information in impression formation. *Journal of Personality and Social Psychology, 18*, 279-284.

Kaplan, M. F., & Kemmerick, G. D. (1974). Juror judgment as information integration: Combining evidential and nonevidential information. *Journal of Personality and Social Psychology, 30*, 493-499.

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences* (2nd Ed.). Belmont, CA: Brooks/Cole.

Klimoski, R., & Mohammed, S. (1995). Team mental model: Construct or metaphor? *Journal of* Management, 20, 433-437.

Luce, R. D. (2010). Behavioral assumptions for a class of utility theories: A program of experiments. *Journal of Risk and Uncertainty, 41*, 19-37.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology, 1*, 1-27.

Myers, J. K. (1979). *Fundamentals of experimental design* (3rd Ed.) Boston: Allyn and Bacon.

Narens, L., & Luce, R. D. (1983). How we may have been misled into believing interpersonal comparisons of utility. *Theory and Decision, 15*, 247–260.

Pennings, T. J. (2003). Do dogs know calculus? *The College Mathematics Journal, 34*, 178-182.

Phillips, L., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology, 72*, 346-354.

Poulton, E. C. (1973). Unwanted range effects from using within-subject experimental designs. *Psychological Bulletin, 80*, 113-131.

Richardson, J. L. (1986). Perspectives on compliance with drug regimens among the elderly. *Journal of Compliance in Health Care, 1*, 33-45.

Rundall, C. S., & Weiss, D. J. (1998). Patients' anticipated compliance: A functional measurement analysis. *Psychology, Health, & Medicine, 3,* 261-274.

Shanteau, J. (1999). Decision making by experts: The GNAHM effect. In J. Shanteau, B. Mellers, & D. Schum (Eds.), *Decision research from Bayes to normative systems: Reflections on the contributions of Ward Edwards* (pp. 105-130). Norwell, MA: Kluwer Academic Publishers.

Shanteau, J. C., & Anderson, N. H. (1969). Test of a conflict model for preference judgment. *Journal of Mathematical Psychology, 6*, 312-325.

Slovic, P., Lichtenstein, S., & Edwards, W. (1965). Boredom-induced changes in preferences among bets. *American Journal of Psychology, 78*, 208-217.

Stevenson, M. K., Busemeyer, J. R., & Naylor, J. C. (1990). Judgment and decision making. In M. D. Dunnette & L. M. Hough (Eds.) *Handbook of Industrial and Organizational Psychology, Vol. 1* (2nd ed., pp. 283-374). Palo Alto, CA: Consulting Psychologists Press.

Tversky, A. (1967). Additivity, utility, and subjective probability. *Journal of Mathematical Psychology, 4*, 175-202.

Tversky, A., & Russo, J. E. (1969). Substitutability and similarity in binary choice. *Journal of Mathematical Psychology, 6*, 1-12.

von Winterfeldt, D., & Edwards, W. (1986). *Decision analysis and behavioral research*. New York: Cambridge University Press.

Weiss, D. J. (1989). Potential methodological contributions of mathematical psychology to patient compliance research. *The Journal of Compliance in Health Care, 4*, 95-100.

Weiss, D. J. (2006). *Analysis of variance and functional measurement: A practical guide*. New York: Oxford University Press.

Weiss, D. J., & Anderson, N. H. (1969). Subjective averaging of length with serial presentation. *Journal of Experimental Psychology, 82*, 52-63.

Weiss, D. J., & Edwards, W. (2005). A mean for all seasons. *Behavior Research Methods,37,* 677-683.

Weiss, D. J., & Shanteau, J. C. (1982). Group-Individual POLYLIN. *Behavior Research Methods & Instrumentation, 14*, 430.

(Manuscript received: 8 July 2013; accepted: 18 December 2013)