15
Order Effects within Personality Measures

Eric S. Knowles, Michelle C. Coker, Deborah A. Cook, Steven R. Diercks, Mary E. Irwin, Edward J. Lundeen, John W. Neville, and Mark E. Sibicky

The Measurement Encounter

As early as 1692, Christian Thomasius had developed 12-point rating scales to measure psychological character (McReynolds & Ludwig, 1984). In the ensuing three centuries, our measurement theories have been refined and have markedly improved the quality of the information that we obtain (F. M. Lord & Novick, 1968; Hulin, Drasgow, & Parsons, 1983).

Most personality measures use the general form where the test maker provides a stimulus to which the test taker provides a response. The stimulus may be a self-description ("I cry easily") that the respondent endorses or disavows, a general statement ("Most things in this world occur by chance") to which the respondent agrees or disagrees, a problem ("Find the picture that doesn't belong") for which the respondent seeks a solution, or even an ambiguous representation (e.g., a TAT picture, a Rorschach inkblot, or an incomplete sentence stem) to which a respondent constructs a story. Current theories of measurement allow test makers to select and develop inquiries that maximize the information that the test maker receives from the inquiry–reply process.

The inquiry–reply measurement strategy engages the respondent in a social interaction with the test maker. Into this social interaction the test maker brings the inquiry and the test taker brings the reply. Out of this interaction the test maker takes information that, with proper scaling and comparison to norms, informs the test maker about the personality of the respondent.

What does the respondent bring out of measurement interaction? At the most general level, a person who engages the test material brings away two things from the encounter:

1. An awareness of and involvement with the issues and constructs employed by the test maker, and
2. A confrontation with self as the respondent attempts to integrate the encounter into the self-concept.

Measurement Reactivity

The question of how the respondent reacts to the measurement encounter has been largely neglected in measurement theories (D. W. Fiske, 1967). When addressed, the question is often framed as a problem of measurement reactivity, defined as "error," and treated primarily as a nuisance for the test maker. Webb, Campbell, Schwartz, and Sechrest (1966), for instance, found that

> The most understated risk to valid interpretation is the error produced by the respondent. Even when he is well intentioned and cooperative, the research subject's knowledge that he is participating in a scholarly search may confound the investigator's data. (p. 13)

Webb et al. (1966) employed a motivational perspective where measurement errors were tied to the self-presentational concerns of the respondent. The respondent, apprehensive about evaluation, was thought to hide the true self or to construct a situationally appropriate (but untrue) identity.

Two things are missed by focusing on measurement anticipation and self-presentation. First, this focus of attention frames issues at a very molar level, using the test as a whole and the entire testing encounter as the units of analysis. Second, the attention on the anticipation and preparation for measurement neglects other interesting issues having to do with how measurement alters the respondent's understanding of the measure or the self.

This chapter focuses on the cognitive rather than the motivational consequences of measurement, in particular on how the measuring process alters the respondent's understanding of the questions and issues addressed in the test. From this perspective, the context effects in which we are interested represent *meaning changes* in the comprehension of the test or of the self and not simply self-presentational strategies. The test item is the appropriate unit of analysis for this inquiry. Specifically, we shall look at how considering one question alters the kinds of answers that are given to the subsequent questions.

Consequences of Being Asked a Question

The cognitive processes initiated by question probes are beginning to be understood, as many chapters in this volume illustrate. Most authors adopt the models proposed by Rogers (1974a) and Tourangeau and Rasinski (1988), who divide the cognitive processes involved in inquiry–reply measurement into a four-stage sequence that includes (a) question interpretation, (b) information/memory retrieval, (c) judgment formation, and (d) response selection. These four stages and the many component processes are described in the other chapters in this volume and in Tourangeau and Rasinski's (1988) thorough review.

From these presentations, it is clear that the task of considering a single question and formulating an answer has many identifiable consequences. Ten of the more important consequences for personality measurement are listed below. These effects are important because they are the sort that may persist and influence answers to subsequent questions.

1. Questions may force an answer to be created where none previously existed (Getzels, 1982; Salancik & Conway, 1975; Sandelands & Larson, 1985)!
2. Respondents construe a question in one particular way, so that one meaning or one interpretation becomes salient and other possible interpretations fade into the background (C. G. Lord, Lepper, & Preston, 1984).
3. Detailed questions may alter the respondents' level of action identification and may make them more susceptible to changing their view of self (Wegner, Vallacher, Kiersted, & Dizadji, 1986).
4. Thinking about an issue tends to polarize the judgments that are made about that issue (Higgins & Rholes, 1978; Sadler & Tesser, 1973; Tesser, 1978; Tesser & Conlee, 1975).
5. Information, memories, and/or attitude structures that are activated by a memory search become more available, more easily accessed, and more influential for subsequent judgments (Bargh & Pratto, 1986; Fazio, Powell, & Herr, 1983; Higgins, King, & Mavin, 1982; Posner, 1978).
6. Declaring an intention or producing an overt answer makes the respondents more committed to their position (Feldman & Lynch, 1988; Kiesler, 1971). Subsequent behavior is then more likely to be consistent with the judgment (Sherman, 1980).
7. A judgment, once rendered, serves as an anchor point against which further considerations may be assimilated or contrasted (Higgins & Lurie, 1983; Strack, Schwarz, & Gschneidinger, 1985).
8. Formulating a judgment and rendering a response themselves activate a post hoc memory search. The search is biased in favor of information that supports the response (Petty & Cacioppo, 1986a). This search can produce entirely new cognitions (Sadler & Tesser, 1973) and can allow existing evidence to be reinterpreted as consistent with the judgment (Tesser & Cowan, 1977).
9. Merely thinking about a complex issue may increase the coherence and interconnectedness of the various facets of the issue (McGuire, 1960; Millar & Tesser, 1986).
10. Considering difficult questions may make respondents develop more complex conceptual structures through which to view issues. For instance, La Rue and Olejnik (1980) found that questions that demanded formal operational thought led young respondents to employ more formal operational responses on a later reasoning test.

These 10 consequences of considering and answering questions suggest the many profound ways that someone's view of an issue and of self may be affected by the measurement process. We need, however, a methodology to study the impact and generality of these influences on personality measurement.

Serial-Position Analysis of Accumulating Effects

Personality tests are particularly useful areas for studying measurement consequences and context effects. A test that inquires about self-descriptions rather than about abstract opinions should engage a deeper and more involved level of processing (Burnkrant & Unnava, 1989; Petty, Rennier, & Cacioppo, 1987). Also, a single-factor personality test should make the consequences identified above accumulate as the respondent considers the same dimension again and again with each additional item. Specifically, (a) respondents should become more confident, committed to, and polarized in their judgment; (b) respondents should become more efficient at making their judgments; (c) respondents should become more consistent and reliable in determining their judgments; and (d) respondents should have a fuller and more organized schema for the construct being measured.

Accumulating reactions have two advantages for studying context effects. First, effects based on a large number of items should be stronger and therefore more evident than would effects based on a single item. Second, the accumulation of reactions should be directly evident as linear trends over the number of items considered. This second property has been particularly important to our research.

The printing press has been a boon to personality research because it has allowed many copies of a questionnaire to be duplicated efficiently. One of the costs of duplication, at least for the study of context effects, is that the printed form confounds the content of an item with its serial position. The first item is always the same, as is the last, and as is every item in between.

This confounding of content with context creates a problem for interpreting item answers. The answer to any particular item on the test includes reactions to the content of that item and reactions that carry over from the previous items. When we inspect answers to the last item on a test, we usually cannot disentangle how much of the answer is due to the content of that particular item and how much is due to the context provided by the previous items.

Of course, interpretation is clouded only at the level of the item answer. Interpretations of the scale score, which is the focus of most personality measurement, is not compromised by the confounding of content with context. However, coming to understand the cognitive processes involved in personality measurement requires a focus at the item level and attention to this confounding.

To investigate the accumulating reactions that people have to thinking about and answering personality test questions, we have had to disentangle the content from the context of the previous items. We have done this by using a randomized latin square to counterbalance item content across the serial positions in a test. For instance, with a 30-item personality measure, we create 30 different forms such that each item (a) appears in each of the 30 serial positions, (b) is preceded and followed by each of the other items approximately half of the time, and (c) is separated by random distances from each of the other items.

Table 15.1 presents a simple randomized latin square for a 7-item measure.

TABLE 15.1. Counterbalance Design for a 7-Item Measure

	Serial Position of Item in Test						
Test Form	1st	2nd	3rd	4th	5th	6th	7th
Form #1	D	A	B	G	F	C	E
Form #2	E	B	C	A	G	D	F
Form #3	F	C	D	B	A	E	G
Form #4	G	D	E	C	B	F	A
Form #5	A	E	F	D	C	G	B
Form #6	B	F	G	E	D	A	C
Form #7	C	G	A	F	E	B	D

Note: The letters "A" through "G" refer to items of different content. Each form presents each item in a different serial position. Over the block of 7 forms, each item (a) appears in each of the seven serial positions, (b) preceeds and follows each other item, and (c) is placed at random distances from each other item. In a sample, equal numbers of subjects receive each form.

The 7 items, labeled A through G, were first placed in a random order, for example, 1st = D, 2nd = A, 3rd = B, 4th = G, etc. This was the order of items used on the first form of the test. For the second form, each letter was increased by one value, except for the highest value (G), which was returned to the lowest (A), for example, 1st = E, 2nd = B, 3rd = C, 4th = A, etc.

This procedure is preferable to a simple rotation scheme in which the last item on Form 1 becomes the first item on Form 2 but the order and distance of items otherwise remain constant. Although our random latin square randomizes these orders and distances, these confounds can be even better controlled by following the prescriptions in Ostrom, Isaac, and McCann (1983). In practice, a microcomputer randomly assigns up to 40 items and then composes and prints multiple copies of each of the counterbalanced forms. We distribute forms to respondents in a random manner but make sure that we have completed replications of the latin square design; that is, with a 30-item measure, we have 30 different forms of the measure, each with a different order of the 30 items, and submit this to some multiple of 30 respondents (e.g., 90, 120, or 150 subjects).

Context Effects within an I-E Test

Counterbalancing item content over serial positions in a test allows the effects of item content to be disentangled from the effects of context. We will illustrate these effects with the results from a data set obtained from 120 respondents to an Internal-External Locus of Control scale (Knowles, 1988, Study 1). The respondents received one of 30 forms composed from W. H. James's (1957) 30-item I-E scale. The answers were prepared in several ways. First, the scores of any negatively worded items were reversed so that positive scores indicated the same end of the scale (Externality). Second, item means and variances were equated by standardizing the answers given to each content item.

Several estimates of context effect are particularly informative.

Mean Shifts

Since the different content items appear in equal proportion in each position, the mean answer at each serial position provides an estimate of the mean test score. If the context created by previous items has no effect, then this mean will remain the same for each serial position. Systematic shifts in this mean, from the beginning to the end of the measure, reflect reactions to the earlier measurement.

The mean answers to James's I-E scale showed no systematic change from beginning to end of the measure, $F(29, 2610) = .86$, NS. Also, the serial position did not interact significantly with item content to affect some items differently from others, $F(812, 2610) = 1.06$, NS. Thus, the mean answers on the I-E test showed no evidence that later answers were affected by earlier answers. We shall show later that (a) some measures do show shifts in mean answers and subjects even in this study systematically polarized their later judgments. For many authors, mean shifts are the only context effects that are measured. Although the I-E test showed no mean shifts, it did show other clear context effects.

Reliability Shifts

The scores at each serial position can be correlated with the sum of the remaining scores. Since all items appear at each serial position, the resulting coefficient is an estimate of the internal consistency of the measure as a whole. If the context created by previous items has no effect, then this correlation will remain the same for each serial position. Systematic shifts in this reliability estimate reflect reactions to earlier measurement.

The reliability estimates did show a significant increasing linear relationship with serial position. Figure 15.1 presents this serial-position effect. The 30 correlations, one for each serial position, were transformed into Fisher's z scores and inspected to make sure that they met the assumptions of a parametric data set. These Fisher z scores themselves were then correlated with the serial position to describe the trend evident in Figure 15.1. The positive linear relationship between serial position and reliability was highly significant, $r(28) = .51$, $p < .01$.

A regression equation ($Y' = .4054 + .0062 \times$ Serial Position) provided a best estimate of the reliability of items in the first position as .390 (Fisher $z = .412$) and of items in the last position as .531 (Fisher $z = .592$). Since the same item content appeared in the first and the last position, this significant increase reflects a reaction to the earlier measurement.

Polarization of Reactions

The total scores on the measure can be used to differentiate respondents into high-, medium-, and low-scoring subgroups. For this analysis, we combined two data sets to obtain 270 subjects (Knowles, 1988, Study 3). Across subgroups, the mean answers to the items at each serial position show a consistent spreading apart. As shown in Figure 15.2, the low-scoring Internal subjects gave mean answers that systematically decreased with the serial position of the item ($r = -.64$). The regression equation estimated Internals' answers to the 1st item to be $Z = -.43$ and answer to the 30th item to be $Z = -.63$. In contrast, the answers given by External subjects were positively but nonsignificantly correlated with serial position ($r = .25$). The regression equation estimated External subjects' answers to increase from $Z = .50$ on the 1st item to $Z = .60$ on the 30th item. Subjects, especially the Internal subjects, became more polarized in their ratings as they thought about and answered more and more of the I-E questions.

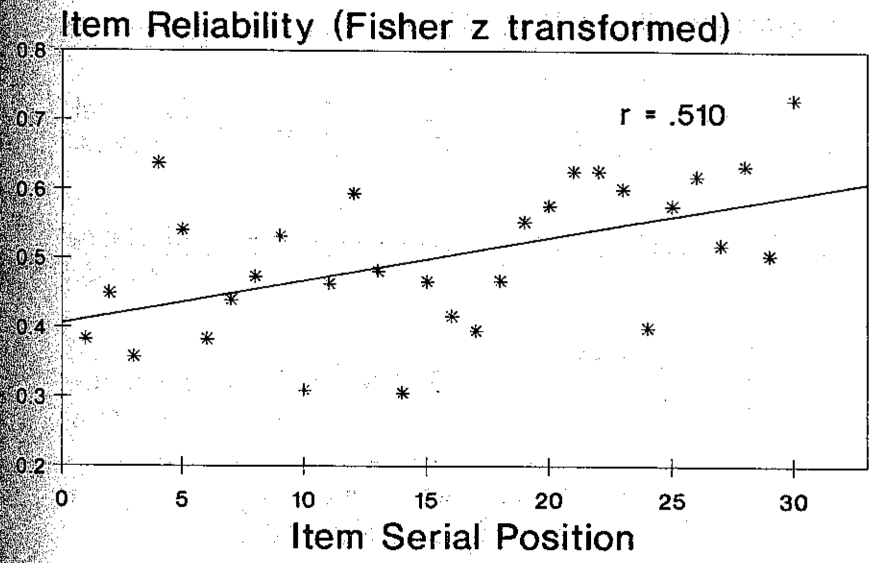

FIGURE 15.1. Reliability Shift on James I-E Scale

Other Consequences of Measurement

We know from other studies that respondents are able to answer later items more quickly and more knowledgeably. In one study (Knowles & Diercks, 1988), 270 respondents were seated at a computer to answer questions from a personality inventory. After familiarizing themselves with the computer procedures by answering a variety of demographic questions, subjects read a screen that described the Personal Reaction Inventory as containing 60 items that concerned a variety of topics, for which there were no right or wrong answers and to which large numbers of people agreed and large numbers disagreed. Following this description, they read and answered 0, 1, 3, 9, or 27 items from either Rokeach's (1956) Dogmatism scale or Taylor's (1953) Manifest Anxiety scale.

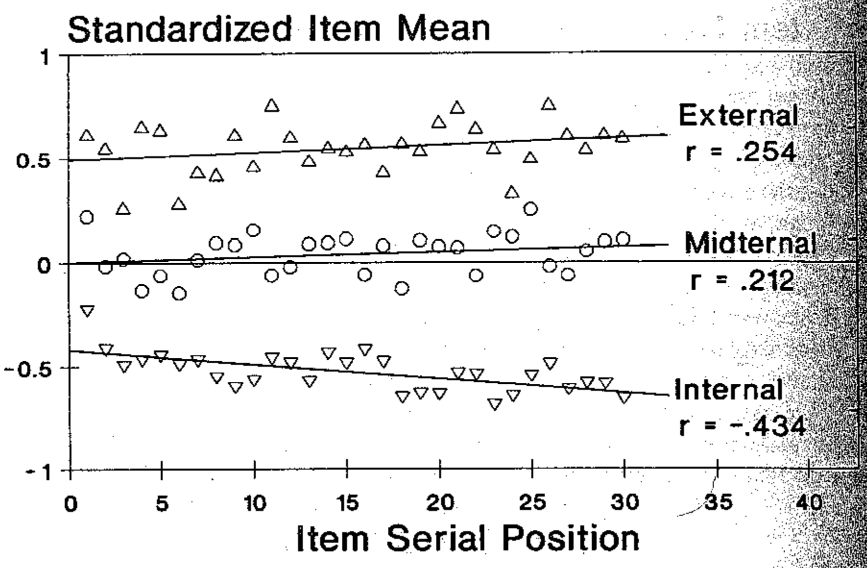

FIGURE 15.2. Polarization of Answers

After finishing this first phase of the study, all subjects were shown eight new items and asked to judge how likely each item was to belong to the Personal Reaction Inventory. The eight items included four that were prototypic of the scale construct and four that were distractor items. In preparation for this experiment, psychology faculty and graduate students rated each of the original test items in terms of how prototypic it was of the test construct. The four items with the highest prototypicality ratings were saved for this second phase of the study and were interspersed with four items from unrelated scales. Respondents entered on the computer keyboard their 9-point ratings of how well the items belonged to the scale. After making these ratings, subjects continued to answer the remaining 27 items from the personality test.

Judgment Accuracy

Exposure to more test items increased subjects' accuracy at recognizing the items that belonged to the scale, $F(4, 260) = 5.46$, $p = .001$, but did not affect their ability to detect the distractor items, $F(4, 260) = 0.60$, $p = .66$. Although anxiety prototypes were judged more accurately than dogmatism prototypes, the effects of exposure to items were identical for the two tests. Figure 15.3 presents the effects of experience on these belongingness judgments, averaged across the two types of personality measures.

Response Time

In this study, we also recorded the response interval from initial display of the personality test item until a response key was pushed. The condition had no main or interactive effect on these response times. As is shown in Figure 15.4, the response times for both personality tests continuously decreased across the 27 items. Of course there are many possible explanations for this decrease in response time. Subjects may have become more efficient judges, may have had the relevant cognitions more available, or may have been lazier. Nonetheless, this study suggests that as subjects answer more and more test items, they answer them both more quickly and more knowledgeably.

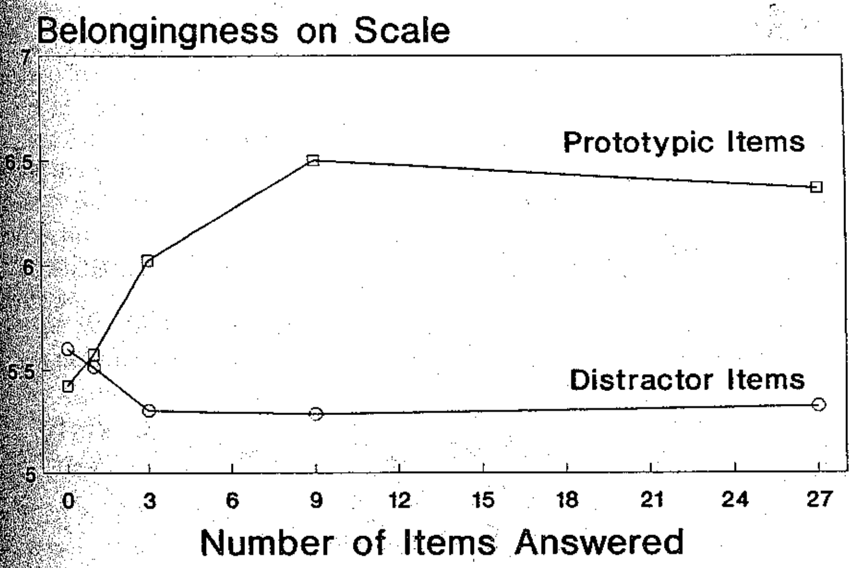

FIGURE 15.3 Learning the Test Construct

Generalizability of Reliability Shifts

Our studies of serial-position shifts in reliability have extended to other measures besides I-E. Knowles (1988) reported equivalent reliability shifts across serial position for measures of I-E (W. H. James, 1957), Dogmatism (Rokeach, 1956), Anxiety (Taylor, 1953), and Social Desirability (Crowne & Marlowe, 1964). In other research we have replicated these findings for I-E (Knowles, Cook, & Neville, 1989a; Knowles, Lundeen, & Irwin, 1988) and Anxiety (Coker & Knowles, 1987; Neville & Knowles, 1990) and extended them to measures of self-acceptance (Knowles, Cook, & Neville, 1989b).

We have not found serial-position changes in item reliability for Snyder's (1974) self-monitoring scale (Knowles, Lundeen, & Irwin, 1988), Beck's (Beck, Rush, Shaw, & Emmery, 1979) Depression Inventory (Knowles, Coker, & Diercks, 1988), or for several extracted MMPI scales (Neville & Knowles, 1990). Although we have not found this context effect universally, we have found it with enough regularity and generality to suspect that it is a widespread phenomenon. We suspect that multifactor measures, where it is more difficult to

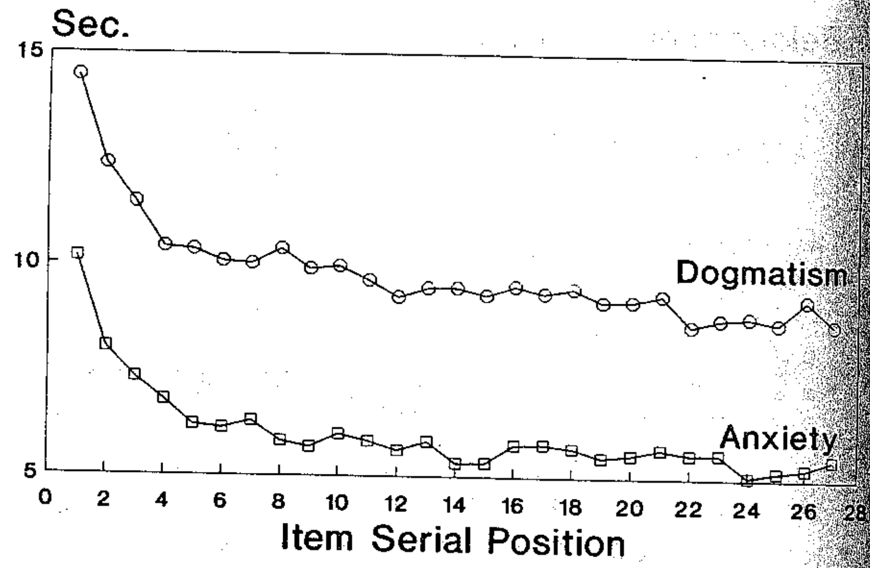

FIGURE 15.4. Response Times to Items

discern a consistent theme to the test, may be less likely to produce the reliability shift (Neville & Knowles, 1990).

Applications to Test–Retest Effects

Retests on a measure have shown two kinds of differences from the original test: reliability shifts and mean answer shifts.

Test–Retest Reliability Shifts

More than 50 years ago, Pintner and Forlano (1938) observed that odd–even reliabilities on several personality measures tended to increase over four testings. Since then many researchers have found that people answer a personality retest more consistently than they answered the first test (D. W. Fiske, 1957; Goldberg, 1978; Howard, 1964; Howard & Diesenhaus, 1965; Schubert & Fiske, 1973; Windle, 1955). Many test developers, including Taylor (1953) for her anxiety measure, report that retests have higher internal consistency than first tests. The greater internal consistency on retest with the same or similar items seems to be a normal consequence of the reliability shifts that we have observed within tests (Coker & Knowles, 1987). The retest elevation in reliability merely perpetuates the changes that occur within the first test.

Test–Retest Mean Answer Shifts

Windle (1954) compiled test–retest data from numerous personality inventories and concluded that a variety of "adjustment" scores showed significantly better adjustment on retest, especially with retest intervals of less than two months. This intriguing mean answer shift on retest continues to be observed for various tests (Chance, 1955; Goldberg, 1978; Payne, 1974; Perkins & Goldberg, 1964; Windle, 1955).

We find that the increased scores on retest are not really a test–retest phenomenon but an item-to-item reaction that is also apparent within the first test. We (Coker & Knowles, 1987) studied two 25-item alternate forms of an anxiety test using the latin square design. The mean answer shift observed within the first testing continued unabated throughout the second testing given a week later, as shown in Figure 15.5.

The shift in mean answers is most often interpreted as an impression management phenomena in which respondents try to present themselves on the retest as more adjusted and socially desirable (Goldberg, 1978; Payne 1974). In Neville, Coker, and Knowles (1988) we interpreted this impression management theory as implying that the subjects who were most anxious would change the most. We divided our Coker and Knowles (1987) sample into high-, medium-, and low-anxiety subgroups, based on the sum of their test and retest scores, and found that each subgroup showed equivalent serial-position decreases in mean answers, both within and between the two test administrations. Since subjects with the most desirable scores changed as much as the subjects with the least desirable scores, the impression management theory did not seem particularly useful. We (Knowles et al., 1989b) recently observed a similar increase in self-acceptance scores within a test, increases that were also equivalent for high-, medium-, and low-scoring subgroups.

Although many personality measures do not show a mean shift within or between test administrations, tests of adjustment, including anxiety and self-acceptance, seem to be susceptible to these context effects. The fact that adjusted and unadjusted respondents show this effect equally is less suggestive of a social desirability explanation and more consistent with a meaning-change interpretation. Answering earlier items may alter the interpretation, recall of relevant information, and meaning of later items in ways that shift the respondent's answer.

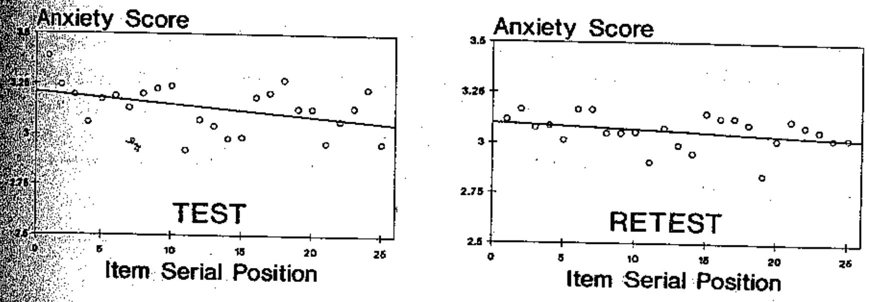

FIGURE 15.5. Mean Answer Shift on Anxiety Test and Retest

Implications for Factor Structure of Measures

One way to conceptualize the meaning-change explanation is that respondents learn through experience to interpret the scale-relevant content of an item and to disregard surplus meanings. Consequently, answers to later items better reflect the scale construct than do earlier items. In Knowles, Lundeen, and Irwin (1988) we used this reasoning and factor analyses to explore the degree to which items saturated the I-E scale construct.

Using counterbalanced orderings of James's (1957) 30-item I-E scale, 15[illegible] subjects were tested and then two weeks later were retested. The item responses were standardized for each item content but left in the order in which they had been answered. In a novel use, factor analysis was employed to identify serial position changes in item saturation.

The scores at each *serial position* were factor analyzed for test and retest. Each serial position included answers from all content items. Since all items contributed equally to the scores at each serial position, the factor analysis of serial positions revealed one large factor representing whatever was measured by all of the items. Figure 15.6 presents the factor loadings of each serial position on the single factor identified at test and retest.

For the first testing, the factor loadings were positively correlated with serial position of the items, $r(28) = .66$, $p < .001$. Based on regression line estimates, items in the first position had an estimated factor loading of .348. The same items appearing at the end of the test were estimated to have a loading of .540 on the same factor.

For the retest, serial position made little difference. All of the factor loadings were high and fairly consistent. However, on this retest, the serial positions had significantly higher average loadings (.495) than did the original test (.444), $t(28) = 2.46$, $p < .05$. The average retest loading (.495) was not significantly different from the estimated loading at the end of the first test (.540), $Z = .73$, NS.

Since every item appears at every serial position, these factor loadings do not represent content differences. Instead, they seem to represent "saturation" differences (Harris, 1975), that is, the degree to which the underlying construct is reflected in the item set. It seems that for the single-factor I-E scale, an item engages less of the test construct when it appears at the beginning of the test than when it appears at the end. Experience with the intervening items clarifies the construct being measured, eliminates surplus meanings from the item, and alters the response characteristics of the item.

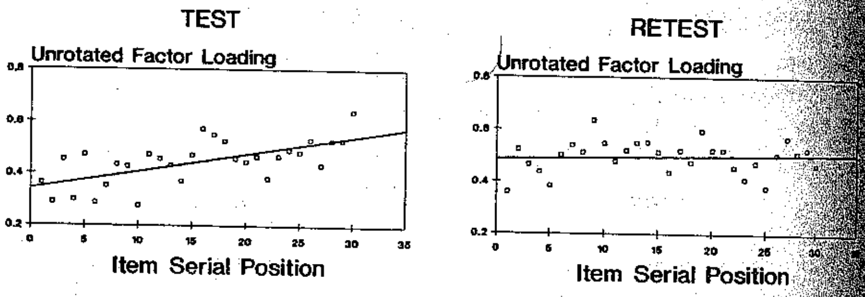

FIGURE 15.6. Saturation of Items at Each Serial Position

Modifiers of the Reliability Shift

Most recently, we have turned our attention to another question: Who shows the reliability shift most strongly? In Knowles et al. (1989a) we investigated the role of cognitive and motivational mediators of the reliability shift. Specifically, we assessed whether people who had a higher Need for Cognition (Cacioppo & Petty, 1982) and higher conceptual ability (as measured by the vocabulary subtest of the Shipley Institute of Living Scale; Zachary, 1986) were more likely to show this serial-position increase in reliability on a 40-item Locus of Control scale.

We found that verbal ability strongly modified the serial-position effect on item reliability, but it was opposite to the one that we had expected. The correlation between serial position and item reliability was $r = .69$ for subjects low in verbal ability but only $r = .18$ for subjects high in verbal ability. This difference in correlations was significant, $Z = 2.85$, $p < .01$.

Figure 15.7 presents the scatter plots and regression lines for these two relationships. Low Verbal Ability subjects showed a much steeper increase in item reliability than did High Verbal Ability subjects. Their later answers apparently profited from the repeated encounters with earlier test items. The High Verbal Ability subjects answered even the first question with a moderately high level of reliability and showed only a slight increase over the 40 serial positions. It seems that their ability to answer a question was barely affected by the previous questions.

We conducted a similar analysis to look for the moderating influence of the Need for Cognition. High Need for Cognition subjects were significantly more internal than Low Need for Cognition subjects $F(1, 398) = 16.79$, $p < .0001$, but showed no difference in the reliability shift, $Z = 0.73$, NS. As shown in Figure 15.8, both subgroups showed the benefits of answering earlier items.

In the Knowles et al. (1989a) study, the subjects' vocabulary ability, more than their cognitive motivation, modified the reliability shift. We assume that the less verbally able subjects answered early items unreliably but, with continued experience, came to discern the test construct more clearly, finally matching or surpassing the reliability that more verbally able subjects exhibited even on the first items. We think that the more verbally able subjects could extract the implicit meaning of the items from the beginning of the test.

Conclusions

This program of research shows that respondents' reactions to tests changed systematically in several ways as they moved from the beginning to the end of

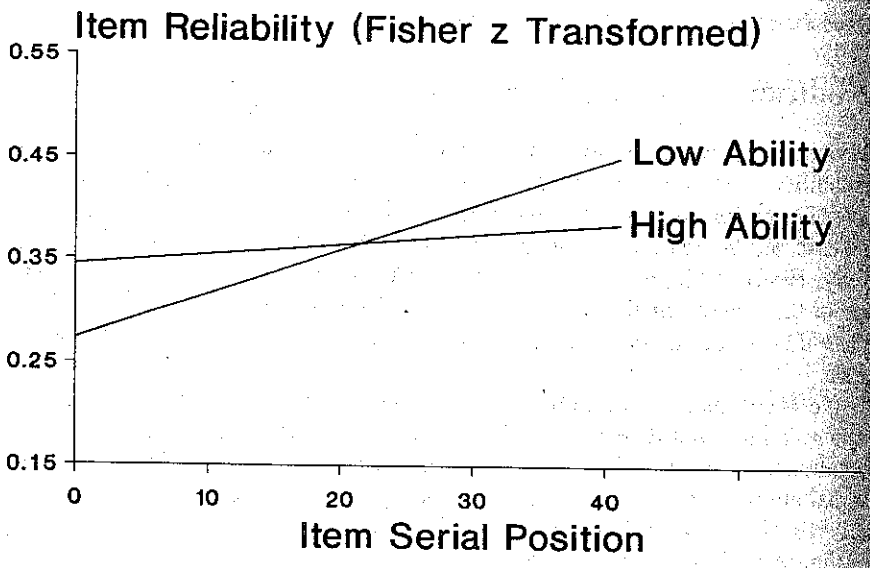

FIGURE 15.7. Verbal Ability and Reliability Shift

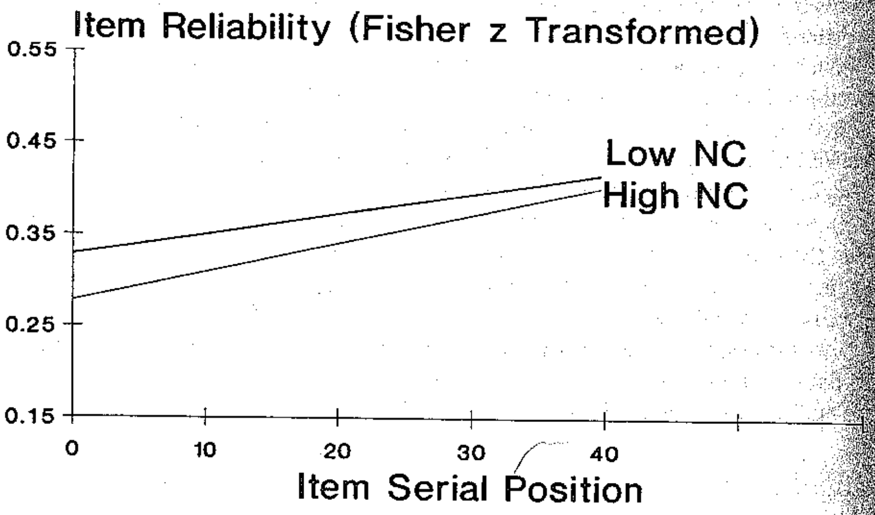

FIGURE 15.8. Need for Cognition and the Reliability Shift

the test and from test to retest. Respondents typically answered questions in more extreme, more consistent, and more reliable ways. These more consistent answers were achieved more quickly and with more understanding of the test construct, particularly by those subjects who would have the most difficulty discerning the implicit meaning of the early items.

Meaning Change

We think that these results are most consistent with what we call the "meaning-change hypothesis." This is an explanation that relies on the cognitive processes involved in considering and answering questions. The four-stage process involved in considering a single item—interpreting its meaning and implications, retrieving relevant information, forming a judgment, and making a response—has a variety of consequences that feed forward to influence the processing of the next item. Consequently, for later items, the construct relevant content is more easily discernible and accessible; construct irrelevant associations can be ignored. Information retrieved and created for earlier items is more available for later items. The early judgments formed for specific items become a generalized judgment concerning the latent construct. Consequently, the response generation process becomes less an information integration task and more the calibration of a generalized judgment to the particular scale values represented by the question and response scale. In short, answers to later items become more informed, more efficient, and more confident.

Alternative Explanations

The meaning-change hypothesis guided this research and, we believe, is most consistent with the results. There are several rival explanations that still need to be considered.

Social desirability is a persistent alternative explanation for instances of measurement reactivity (Goldberg, 1978; Webb, Campbell, Schwartz, Sechrest, & Grove, 1981). Social desirability does not seem to offer a reasonable alternative explanation for the reliability shifts, response polarization, or construct learning effects that we have found. However, we believe that social desirability plays a part in the mean shifts that have been found for tests of adjustment (e.g., Windle, 1954). Social desirability is likely to influence (a) the content of the thoughts that are retrieved in response to each question, (b) the arguing and counterarguing that occurs during and after considering a question, and (c) the ease with which questions and thoughts are assimilated into the self-concept. Although there certainly are situations where people decide to "fake good" or "fake bad," we believe that the more general influence of social desirability is fairly subtle and has its impact early in the response process.

Cognitive laziness is another possible rival explanation (Israelski & Lenoble, 1982; Krosnick, in press). In this view, respondents become fatigued or bored with repeatedly considering similar questions and simplify the strategy that they employ to formulate answers. Rather than fully considering and integrating their information about questions, respondents come to adopt a strategy that "satisfices"; that is, one that produces, with minimal effort, a just adequate rather than optimal response. Krosnick (in press) provides a careful and provocative review of how satisficing strategies might be employed in answering questions.

We have concluded that respondents become more efficient in answering later personality items and are able to neglect the irrelevant meanings of an item to concentrate more directly on the abstracted test construct. These effects, seen through a cynical eye, may have the look of laziness.

The meaning-change and satisficing alternatives involve more than whether one is looking at the same phenomenon through pollyanna or jaundiced eyes. Several tests between these alternative explanations are possible. One test would look at the effects of ego involvement. Ego involvement should increase the motivation to process the items and should reduce the reliability shift, but in different ways according to the meaning-change and satisficing hypotheses. Under meaning change, increased motivation should raise the reliability of early-appearing items; but, if the satisficing hypothesis is correct, motivation should reduce the reliability of later-appearing items. A second test would look at the effects of task demands or distraction. Increasing the cognitive task load should reduce the reliability shift under the meaning-change explanation but should increase the reliability shift under the satisficing theory (Krosnick, in press).

Ambiguity of Measurement

These studies have an intriguing implication: They bring the Heisenberg (1958) principle to personality measurement. The act of measurement alters the meaning of the measure as the respondent becomes educated about self and about the test construct. When these changes occur at the item-to-item level, it is very difficult to tease apart how much of the answer reflects the personality before measurement and how much reflects an interaction with the measurement.

Measurement that engages the respondent in an inquiry–reply encounter is not "objective" in the sense of being outside of and apart from the respondent. It is a subjective and interactive process, in which both the respondent and the test taker bring something into the encounter and both take something away from it.

16 Context Influences on the Meaning of Work

Gerald R. Salancik and Julianne F. Brand

Order effects in surveys are a special class of context effects on opinions. In this chapter, we examine context effects on a job reaction survey and show that context can be manipulated to study how individuals use information about their job experiences to generate job descriptions.

Researchers who study job reactions propose that individuals are more or less motivated to do their work and more or less satisfied with their jobs when those jobs possess certain features. In particular, a job is said to be motivating (or satisfying for the individual) if the work is organized to allow the individual *autonomy* and discretion as to how to do it, if it lets the individual use a *variety* of his or her talents and abilities, if it permits the individual *identification* with the outcome, if the individual can see the *significance* of his or her work for others (customers, co-workers), and if the individual gets *feedback* from the work about how well or poorly it is going. The general relationship between job reactions and these characteristics has been well established. Jobs characterized by autonomy, identification, variety, significance, and feedback are typically rated by those who do them as more satisfying and more motivating (Hackman & Oldham, 1980).

Less is known about how individuals characterize their work along these dimensions. The authors of one instrument for describing jobs, the Job Diagnostic Survey (JDS), used in several hundred studies, assume that workers' responses are simply reports on objective features of their work (Hackman & Oldham, 1980); that is, the individuals are reporting on what they know about their jobs. One indication that this might be the case is that observers will often describe jobs in the same way as they do their occupants. Other research, however, has shown that job descriptions are affected by social suggestions (e.g., Griffin, 1983; O'Reilly & Caldwell, 1979) and various response and method artifacts (Glick, Jenkins, & Gupta, 1986; Idazak & Drasgow, 1987).

Consistent with both positions, we argue that survey instruments asking persons to characterize their personal experiences along a set of dimensions—such as the Job Diagnostic Survey—are influenced necessarily by both content and context. We assume this because such instruments, by design, ask persons to translate their experiences into specified descriptions. If the task is taken seriously, then in fulfilling its request they would review their knowledge of the situation and derive interpretations meaningful for that context (Salancik, 1982). Respondents, of course, may repeat descriptions given to previous inquiries, but their earlier views will reflect the knowledge and interpretations then available. Whenever views are generated, anything affecting the recall or use of information about the situation will affect judgments.

Theoretical Framework

There are at least four influences of context on responses. One arises if a context directly primes a response, as is common when a person acknowledges how nice another looks before going on a date. Another occurs when a context makes information salient, which in turn affects a response, as might happen when a person going out the door with a date is prompted to remember whether the gas tank is filled. These two effects are commonplace biases that are also easily detected because they affect most respondents in the same way. Two additional effects, however, are conditioned by a person's own experiences or knowledge. One effect comes from the differential relevance that a person attaches to knowledge in different contexts even when it is readily available in each. In evaluating a person in the context of a date, one might consider the knowledge that the person wears glasses as important information for the evaluation but might easily overlook such trivia when evaluating the person in the context of a marriage. A related effect arises from the differing implications derived from a particular knowledge rather than from its relevance. Knowledge that a person wears glasses may entail either negative or positive implications in different circumstances.

These latter influences are closely related to the kinds of order effects noted for survey questions. As an illustration, consider a person with considerable knowledge about his preferences and experiences with several fruit juices. This person really loves orange juice, easily recalls its taste and aroma, and readily sees himself reaching for it on entering a grocery store. In contrast, he has only a mild regard for grapefruit juice and fleeting experiences. Finally, he has a definite dislike for apple juice, quivers at its pungent odor and the bland, flat, untextured residue that it leaves in his mouth as it is swallowed. Suppose asking this person two questions: "Do you like orange juice?" and "Do you like fruit juices?" We might expect such a person to have a lower probability of answering "yes" to the second question if it were asked after the first than if it were asked before. We would expect this because having confessed his passion for orange juice to the first question, a repetition might seem a bit out of place for the second, especially since it might imply he likes all juices when he clearly does not. If asked about fruit juices first, however, he might surely wish to speak out strongly for his favorite. For this individual, context differs from one order to another and would be revealed by comparing them.

Yet, because different individuals have different experiences and knowledge, the order-effects paradigm will be limited for revealing such effects as a general rule. A person with the opposite preferences would not be similarly affected. Hence, we may be unable to observe real context differences when they are present simply by seeing how different orders of questions influence responses. Equally important, we shall be unable to learn how individuals generate their opinions in the first place.

To study context effects that are conditioned by personal knowledge requires a different kind of paradigm. The one that we use here is the context-priming paradigm. In this paradigm (Salancik, 1982), an investigator assumes (or knows) that certain knowledge is in the domain used by individuals in coming to their opinions. To study how opinions are generated and how context affects them, one varies the context in some way that is theoretically relevant to hypotheses about the opinions under study. For example, in the theory of job reactions, job characteristics are assumed to be stable properties of the way in which a person's work is organized and directed in an organization. If a person is free to do a job in whatever way he or she deems appropriate, he or she should attribute autonomy to it. If the job directly and visibly affects the welfare of others, it should be invested with great significance. As former Governor Jerry Brown of California said of university professors, "They should be paying the State for the privilege of having such wonderful jobs," alluding to the possibility that professors can do whatever they like whenever they like and have a great time playing in their research labs while admiring students writing down their every word. Yet, in this job, as in others, there are contradictions. Passing gems of wisdom every day can consume energy and time far in excess of the doting affirmations received. Such consumptions could clearly diminish a sense of significance. We expect that these and other feelings would depend severely on whether one thought about teaching in a context of oneself or one's students. From the point of view of theories about job reactions, such an outcome would question assumptions that jobs have stable properties that spawn motivations and satisfactions.

Thus, our purpose for the study described here was to evaluate how opinions about jobs are affected both by content and by context. In brief, workers were queried about their job content, primed by one of two context cues, and then asked to characterize their jobs on the Job Diagnostic Survey (Hackman & Oldham, 1980).

Overview of Study

The jobs studied were the teaching jobs of University of Illinois graduate assistants, which vary widely in the instructional, administrative, evaluative, counseling, and social activities involved. To ensure that all respondents were aware of their jobs' content, we first gave them a list of teaching assistant (TA) activities and ask them to place a checkmark by the ones that applied to their jobs.

This also gave us specific knowledge about the tasks comprising their jobs. An underlying assumption was that individuals will process knowledge of their job activities to derive meaning but that the meaning derived will vary with context. The design was constructed to allow us to evaluate the alternative influences that context might have on job descriptions. If the context affected relevance or meanings, then job activities would associate with JDS characteristics differently in different contexts. If, on the other hand, task content directly implied certain job descriptions, then the same job activities would imply the same characteristics regardless of context. And if job interpretations differed regardless of job activities, then context affected either information recall or response biases.

Context was manipulated by focusing the TAs' attention on themselves or on their students, a choice based on pilot interviews and our own experiences. Interviews suggested that the TAs held their jobs in contradictory regards. They vacillated between thinking of the job's effects on their students and on their own lives and obligations. Thus, at times, teaching was viewed as a good opportunity to gain experience in organizing and integrating knowledge from one's field, but at other times, it was a drag that kept one away from academic goals. These variations in meaning from the same content were consistent with our theoretical views.

Method

Forty graduate teaching assistants at the University of Illinois were interviewed about their jobs. The individuals were sampled randomly from the University's list of graduate TAs stratified to cover major branches of studies—the physical, biological, and social sciences, the humanities, the performing arts, and engineering. All were graduate students employed half time. TA jobs differ greatly, ranging from independently teaching entire undergraduate courses, through leading discussion sections of a large course, to assisting in setting up demonstrations for a professor-taught course. Interviews were arranged by contacting the TAs and meeting them at convenient locations on campus, their places of work or their homes. The interviewer was a female graduate student who also taught undergraduates. Each interview lasted one to three hours. Respondents had no knowledge of the study before being solicited and interviewed.

Procedures

The study was introduced to respondents in a straightforward manner:

> We are doing this study to find out more about what graduate teaching assistants at the University do in their jobs, how the jobs of instructors in different departments vary, and how instructors feel about their jobs.

Interviewing was in three stages. First on a list of 95 activities each TA checked off those that were typically and regularly part of his or her job, then answered a question that manipulated context, and finally assessed the job on the JDS and evaluated his or her job satisfaction.

Job Activities

The list of 95 activities was constructed during a two-month pilot study of a sample of 15 University teaching assistants. The individuals free associated about the activities that made up their jobs. Several hundred items resulted. These were edited to incorporate similar activities under the same description. Editing was repeated until the final 95 were settled on. For ease of presentation, the items were grouped into five general categories: (a) instructional activities, (b) counseling activities, (c) administrative activities, (d) maintenance activities, and (e) evaluative activities. Four examples of the most common activities in each are presented in Table 16.1.

To collect information from each respondent about the activities comprising his or her job, the interviewer handed out the list of activities and asked the respondent to read each item and judge if the indicated activity was something done as a regular or normal part of the job. It was explained that "regular" meant "the activity is something you normally do in your particular job even if it is done only once during a semester." Illustrations were then provided indicating that an activity such as "grading final exams" might only be done once in a semester but was a regular part of the job if done each term. It was also explained that our interest was in the job as it was actually done rather than as others might do it or as the University might have defined it. The TA was then left alone to check off the activities, and the interviewer sat nearby to answer questions. Questions were minimal. After marking the activities list, the respondent was asked if any items were unclear or puzzling. The purpose of this question and the checklist was to ensure that the respondent's job activities were salient and accessible from memory.

Context-Saliency Manipulation

After reviewing the activity list, the interviewer injected the context manipulations with

> Now, we'd like to ask you about some of the effects that your job has had. Specifically, we would like to know how you do your job, what you do in your job, how this benefits (you/your students).

The parenthetical variation in the last sentence constituted the context manipulation. The respondent was asked to think about such effects and describe them. Reports lasted about three to five minutes, and the interviewer took notes but said nothing during them except affirmations ("Uh-uh" or "I see"). Conditions were assigned by alternating interview schedules between conditions until about 20 respondents were available for the "self" and "student" contexts. Each condition served as a control for the other.

TABLE 16.1. Illustrative Activities Checked by TAs, Grouped by Category

1a. Instructional Activities
- Eliciting questions for class discussion
- Summarizing readings to the class
- Paraphrasing student comments in class discussion
- Skimming media for class relevant ideas or materials (e.g., TV)

1b. Counseling Activities
- Offering course advice
- Counseling students about academic matters outside of your course of program study
- Giving advice on personal development, growth, or time management

2a. Administrative Activities
- Ordering books through publisher or University system
- Putting readings on reserve in library or elsewhere
- Ordering lab or classroom supplies (e.g., chemicals, paper, paints)
- Keeping attendance records

2b. Maintenance Activities
- Attending ongoing seminars for TAs held by the course supervisor
- Consulting with administration about class management issues
- Selecting wardrobe for class or laboratory presentations
- Going for drinks with those associated with class

3. Evaluative Activities
- Defining grade criteria
- Scoring the homework problems or exercises
- Providing critiques or constructive comments of the quality of work to students
- Preparing test or quiz questions

Job Characteristics

Following the context manipulation, respondents described their jobs according to the JDS scales of Hackman and Oldham (1980). The job characteristics part of the survey included five questions that directly asked the respondent to assess the extent to which the job possessed autonomy, identification with the work, variety in the use of skills, significance or importance of the work for the well-being of others, and feedback from the work itself about performance. These assessments were requested on 7-point scales with appropriate verbal descriptions for the ends and middle of the scales. In addition, respondents marked on a scale from 1 to 7 how accurately each of 14 sentences described their jobs. Skill variety was assessed in the first case with "How much *variety* is there in your job? That is, to what extent does your job require you to do many different things at work, using a variety of your skills and talents?" and in the second case by "The job requires me to use a number of complex or high-level skills" and "The job is quite simple and repetitive." All scales were used, aggregated, and scored as recommended by their authors.

Job Satisfaction and Motivation

Respondents also answered the job satisfaction questions asked in the JDS procedure, which ask about several facets (see Table 16.2). It was expected that job satisfaction judgments would follow from the job characteristics generated during the job description phase of the study. As such, the manipulated context should have affected job satisfaction if it affected job attributions.

Analyses and Results

Means of the various scales used in the study for each context are presented in Table 16.2. Contexts did not differ by the number of reported job activities, suggesting that they were not accidentally confounded with job content reports. However, TAs induced to think of their jobs in terms of themselves reported less "internal motivation" ($t=3.13$), more "significance" ($t=2.10$), and less "feedback" from the job ($t=2.16$). These effects, we shall see, were primarily the result of context-manipulation effects on the TAs' use of job activity information to infer job features. As required by our assumption that job characteristics are related to knowledge about job experiences, a significant canonical correlation (.578) was observed between the set of job activities and the set of job characteristics (chi-square $=42.97$, $df=25$, $p=.014$).

Our main analysis concerned the effects of context on the relationship of job activities to the characteristics that respondents attributed to their jobs. To test whether JDS characterizations were the result of the content of the job activities alone, the number of activities reported in each category was regressed against the JDS categories. To test whether the context induced TAs to use the information about their job activities differentially, interaction terms were regressed in a similar manner. Interaction terms were constructed by coding context subgroups +1 and −1 and multiplying these codes with the job activities of each respondent.

Job activities were assessed as the number of items that the individual checked summed across categories of instructional, counseling, administrative, maintenance, and evaluative activities. The first two were added together to form a category of "educational" activities, and the second two were added to form a category of "maintenance" activities. This grouping reduced the activity categories to a smaller but logically meaningful set so that stable regression analyses could be performed.

Test of the argument that context affects how respondents interpret their jobs, given their activities, involves comparing regression coefficients for the activities independently of context and when they interact with context. If interaction terms are significantly associated with JDS reactions, we can conclude that the context affected respondents' use of job activity knowledge in forming opinions about their jobs. If only the coefficients for activity reports are significant, it indicates that the respondents were affected only by their job experiences.

TABLE 16.2. Means for Self/Student Context Groups on Job Activities, Job Characteristics, and Attitudinal Measures

	Context		*p*-Level
	Self ($n = 21$)	Student ($n = 19$)	Difference
Reported Job Activities			
Instructional	10.57	12.84	
Counseling	1.42	1.00	
Administrative	12.33	13.00	
Maintenance	7.81	7.84	
Evaluative	12.57	13.05	
Reported Job Characteristics			
Autonomy	3.75	4.02	
Identification	4.37	4.67	
Variety	4.64	4.46	
Significance	4.22	3.89	.05
Feedback	4.35	4.65	.05
Attitudinal Measures			
Internal motivation	5.58	6.21	.01
General satisfaction	5.43	5.60	
Job satisfaction	5.05	5.29	
Security satisfaction	4.81	4.74	
Compensation satisfaction	3.91	4.34	
Supervisor satisfaction	4.17	4.42	

Table 16.3 presents the unstandardized regression coefficients predicting JDS characteristics from the activities in each category and their interactions with the self/student context. Since the interaction terms were uncorrelated with job activities (the nine *r*'s range between +.15 and –.05), a stepwise analysis was done with variables selected according to their significant contributions in reducing variance. In the case of every JDS characteristic except "autonomy," one or more interaction terms entered the regression first and was significantly associated with the characteristic.

Three of the 10 significant coefficients in Table 16.3 relate activities with job characteristics independent of context. Respondents attributed more "autonomy" to their jobs when the job involved them more in instructional and less in maintenance activities, and they attributed less "significance" to their work when they were involved in evaluation.

Seven of the 10 coefficients relate activities with job characteristics dependent on context. These indicate respondents from each condition reacted differently in forming interpretations about their jobs, given their similar experiences. Since the probability was only .14 ($1 - .95^3$) that at least one of the three activities would have been significantly related to a job characteristic, the fact that four of the five characteristics had significant interactions suggests that context had a powerful effect on the meanings that the TAs derived from their experiences.

TABLE 16.3. Significant Coefficients from Stepwise Regression of Job Activities and Their Interaction with Context (IC), on Job Characteristics

	Job Characteristics				
Job Activities	Autonomy	Identification	Variety	Significance	Feedback
Educational	.112**	—[a]	—	—	—
Maintenance	–.071**	—	—	—	—
Evaluative	—	—	—	–.058*	—
IC-EDUC	—	—	—	–.098**	—
IC-MAIN	—	–.025**	.035**	.035*	—
IC-EVAL	—	—	–.050**	.052*	–.041*
R^2	.27	.25	.19	.38	.13

[a] — = NS (not significant).
*$p < .05$. **$p < .01$.

Although the results in Table 16.3 tell us that the context subgroups differed in the way in which they formed job opinions, they do not tell us exactly what meanings each subgroup derived from their job experiences. To assess these, we regressed job characteristics simultaneously on context, the job activities reported for each area (educational, maintenance, and evaluative), and the interaction of context with each activity category. The equation estimated is

$$JC = b_0 + b_1 I + b_2 M + b_3 E + b_3 C + b_4 I{*}C + b_5 M{*}C + b_6 E{*}C + e,$$

where JC is the job characteristic; C is the context coded 1 or –1; and I, M, and E are the numbers of instructional, maintenance, and evaluative job activities reported. The overall effect of activities on the characteristics attributed to the job is determined by their partial derivatives, which

$$\text{Educational effect} = b_1 + b_4{*}C;$$
$$\text{Maintenance effect} = b_2 + b_5{*}C;$$
$$\text{Evaluative effect} = b_3 + b_6{*}C.$$

Note that the effect of each job activity is moderated by context. Since the contexts were coded +1 and –1, the overall effect of an activity in the self context is the sum of the coefficients; in the student context, it is their difference. We did this analysis for each of the job characteristics. The statistically significant results are summarized below.

TAs in the self condition who reported that their jobs involved educational activities saw those jobs as having autonomy but lacking significance, whereas those in the student condition who reported doing educational activities saw their jobs as having both autonomy and significance. The respondents from the self context who reported doing maintenance activities saw their jobs as lacking identification, whereas those reporting maintenance activities from the student context saw their jobs as lacking autonomy and variety. The respondents from the self context who reported evaluative activities described their jobs as lacking variety, whereas those reporting evaluative activities from the student context saw their work as providing feedback but detracting from identification and significance.

In short, context seemed to play a major role in determining the meanings that respondents derived from their work experiences. As an indication of its relative importance, Table 16.4 presents the amount of variance associated with the job activities alone and the incremental amounts associated with context and its interaction with activities. The results are from hierarchial regressions of each job characteristic, constrained to enter the main effects first and interactions last. Although this evaluation is conservative, the results reinforce our interpretation of Table 16.3. Only in the case of "autonomy" did job activities by themselves account for a significant amount of variance. Although the context manipulation appears to have had a direct and independent effect on "feedback," most of its effects were through interactions with job activities. Significant increments to explained variance were contributed by interaction effects for respondents' views on "significance," "variety," and "identification."

The results from Tables 16.3 and 16.4 are particularly striking because responses to the five job descriptors did not themselves correlate much for this sample; only "autonomy" and "feedback" correlated significantly ($r = .386$). Correlations for the other nine pairs of job characteristics ranged in absolute value between .006 and .271.

A final indication that context affected respondents' interpretations of their work experiences comes from analyses of the satisfaction measures. TAs who were induced to think of the job in terms of themselves reported less "internal motivation." However, this effect was mainly due to the differential importance of educational and maintenance activities for these respondents. They reported less "internal motivation" when they were more involved in maintenance activities ($b = -.52$; $p < .01$). Recall that these subjects also reported feeling less identified with their jobs when they were involved in maintenance. If we control for these effects, the direct association of context with "internal motivation" vanishes.

Discussion and Conclusions

The data are clear in suggesting that context had a strong effect on the way in which TAs derived meaning about their teaching jobs from their experiences. The data are also clear in suggesting that knowledge about the content of their jobs was insufficient as a basis for their interpretation of its features. Using the TAs' reports of their specific work activities as indicators of the content of their work experiences, we found that these activities bore little systematic direct relationship to the TAs' descriptions of their jobs on the Job Diagnostic Survey. Yet this lack of relationship is not because the job activities were irrelevant for their job descriptions. The reported activities were, in fact, systematically related to job descriptions, but the particular relationships were dependent on the context primed for a respondent. When the TAs were primed to think of their work in terms of their students, they tended to derive positive meanings from their instructional and evaluative activities and negative meanings from their administrative duties. The TAs primed to think of their work in terms of themselves derived negative meanings from each area of work activity.

TABLE 16.4. Incremental Variance Results from Forced Hierarchial Regression of Job Activities and Context on Job Characteristics

Job Characteristics	(1)	(2)	(3)	Overall R^2
Autonomy	.30*	.02	.03	.352**
Identification	.09	.03	.22*	.337**
Variety	.02	.02	.15*	.225
Significance	.07	.07	.35*	.490**
Feedback	.03	.14*	.11	.285
df	3/36	1/35	3/32	7/32

*Indicates significant ($p < .05$) increment in explained variance over previous step.
**Indicates significant ($p < .05$) overall R^2.
Step (1)—Adds job activities content.
Step (2)—Adds "self/student" context.
Step (3)—Adds interaction of context × content.

Interpreting these results is helped by the unique design used in this study, the context-saliency paradigm (Salancik, 1982). In this design, context differences that are believed to influence the knowledge that individuals use in forming their opinions or how they use that knowledge are manipulated independently of the knowledge. Such a design allows a very explicit evaluation of context effects and enables one to determine if context has a general biasing effect on opinions or influences the way in which individuals form opinions from the knowledge available to them. For the present study of TAs, for instance, it is clear that the context manipulations primarily affected the relevance that respondents attached to their various job experiences. Overall, seven of the coefficients relating job activities to job descriptions were significant (alpha set to .05) for the student context, whereas only four were significant in the self context. Moreover, three of the seven coefficients for TAs primed to think about their students related to their evaluative activities, indicating that these activities were very relevant for deriving the meaning of their work. Finally, in only two cases did the two context groups relate the same activity to the same job descriptions, and only for one of these were the signs of the coefficients opposite. In short, context seemed primarily to have affected what knowledge the TAs used in forming opinions rather than the implications that they derived from that knowledge.

Acknowledgments. We thank the Office of Instructional Resources, University of Illinois at Urbana-Champaign, for help in various stages of the study, and Greg Oldham, Joe Porac, Carol Kulik, and Ellie Weldon for comments on earlier drafts.

17
The Psychometrics of Order Effects

Abigail T. Panter, Jeffrey S. Tanaka, and Tracy R. Wellens

The chapters in this volume and its predecessor (Hippler, Schwarz, & Sudman, 1987) attest to a broad interest in item-order effects. A better understanding of their causes and implications has come from perspectives in cognitive psychology, social cognition, and survey methodology. The importance of contemporary psychometric theory for evaluating order effects has been overlooked. We believe that recognition of psychometric contributions can help researchers better understand item-order effects. One obstacle faced thus far has been that many of the latest psychometric developments have not been readily accessible (or comprehensible) to researchers in this area. The goal of this chapter is to show how such contributions logically relate to testing and interpreting item-order effects. In doing this, we define some necessary preconditions that should exist for the effects obtained to date, and for those obtained in future research, to be interpreted unambiguously.

In the first section, we introduce and review some basic principles of modern psychometric theory that are relevant in understanding item-level response processes. More specifically, we consider item-order effects as they occur within a single construct, as well as across constructs. Throughout these discussions, we illustrate our thinking with examples, including the General Social Survey (GSS) abortion items. Finally, we discuss how psychometric theory allows researchers to specify the characteristics of items, thus better informing research on item-order effects.

The Relevance of Psychometric Theory for Understanding Item-Order Effects

Consider a simple examination of an item-order effect. Two items, A and B, are presented to a respondent in either order AB or order BA. An effect is considered to be present if the response to item A or the association between items A and B is changed in some way by the order manipulation. Built into this strategy are certain assumptions about the nature of the items, A and B, and the nature of the construct (or constructs) that are presumed to underlie these items. More concretely, this order manipulation can be illustrated in survey research employing the classic Schuman and Presser paradigm (e.g., Schuman, Presser, & Ludwig, 1981). Distinctions are made between "general" items, which are thought to assess a respondent's broad underlying attitude, and "specific" items. The specific item can either represent an item drawn from the same construct as the general item or from a different, but potentially associated, construct. Researchers select specific items to tap an ostensibly narrower context than general items.

One relatively robust effect reported in this paradigm occurs when a general item is juxtaposed with a specific item. This presentation creates an item-order effect (as defined above), where the proportion of respondents who agree with a general item is changed (reduced) when the specific item is presented first. One prominent explanation of this effect draws from communication rules and states that the findings are compatible with given-new contracts in information processing (e.g., Strack & Martin, 1987).

What has been generally missing from this research is a psychometric perspective. Measurement issues often play handmaiden to the "more interesting" substantive questions of this research. An investigator may carefully consider what construct will be measured, whereas comparatively less effort may be devoted to the determination of the multiple ways in which a construct can be measured. This approach often results in measures with unknown reliability and an ambiguous relation to the underlying constructs of interest. The selection of measures is typically guided by pragmatics, the history of a research literature, or investigator assumptions and "intuitions."

We explicate later in this chapter the psychometric assumptions that must hold true when inferring order effects based on comparisons between items. To interpret an item-order effect, a number of preconditions must exist within the data. When these assumptions are not met, the interpretation of these empirical findings is obscured, since there is no formal evaluation of how the units of analysis (i.e., items) relate to their underlying hypothesized construct(s). We begin, however, with an overview of developments relevant to the understanding of order effects from a psychometric perspective.

Historical Psychometrics Perspectives

As reviewed by Leary and Dorans (1985), there has been a relatively extensive history of research on item-rearrangement effects in the educational measurement literature. Adopting their structure, the study of item-order effects can be classified into three periods of research emphasis. In the first period, which parallels the initial interest in order effects in the survey methodology literature, the presence of order effects when individuals read and respond to self-report items was simply acknowledged. The key studies during this period investigated item-order effects by manipulating order in three conditions: (a) easy-to-difficult, (b) difficult-to-easy, and (c) random (cf. Mollenkopf, 1950; MacNicol, 1956). These different presentation orders can be thought to be analogous to the general/specific–specific/general item orderings, although the item content and assessment goals in the educational literature clearly are different.

The second phase of the rearrangement research began to address more process-related issues in interpreting item-order effects. Work on this domain with educational tests considered individual difference variables such as test anxiety or achievement level as possible moderators of observed item-order effects (e.g., Hambleton & Traub, 1974; Smouse & Munz, 1968). For example, an individual high in anxiety might show different patterns of item response when difficult items were presented first followed by progressively easier items than the converse. The experimental manipulation of item order or response context represents another way to understand process issues (e.g., Tourangeau & Rasinski, 1988). In the latter case, focus is directed toward characteristics of the response situation and away from properties of the individual.

The final stage outlined by Leary and Dorans (1985) uses modern psychometric theory in conjunction with psychological process theories to characterize the mechanisms underlying order effects (see also Whitely & Dawis, 1976; Yen, 1980). From our perspective, we concur with the work of Leary and Dorans in their suggestion that a more appropriate emphasis in understanding these effects may not be the single-item juxtapositions but rather the rearrangement of intact test subsections (or "testlets"), each reflecting different item characteristics (e.g., item specificity). Drawing on logic that we have developed more fully elsewhere (Tanaka, Panter, Winborne, & Huba, 1990), we argue that the availability of multiple indicators to demonstrate order effects is preferable to the more typical single-item approaches. This focus concentrates on comparisons at the construct level.

Items and Their Relations to Underlying Construct(s)

If responses at the item level are to be compared, then it is of primary importance at the outset to understand the relation of target items to their respective underlying construct(s). In certain research contexts, the investigator may be comparing the item order of two items from the same construct. This emphasis often is seen in educational or personality assessment, where interest may be in a single underlying attribute (e.g., ability or trait). In this case, item-order effects would occur within a given construct. The work of Knowles and his colleagues (Knowles, 1988; Knowles et al., chap. 15, this volume) exemplifies this approach. Alternatively, the two items being evaluated may be hypothesized to represent two different, but potentially related, constructs. This emphasis is more characteristic of survey methodology applications, where order effects are likely to occur between constructs.[1]

[1] We thank Norbert Schwarz for making this distinction.

Although empirical tests regarding the relation of the target items to their respective constructs are possible, they are rarely conducted. Instead, investigators tend simply to accept the assumption that the items are assessing the constructs that they think they are, to the degree that they think they are, because the items superficially reflect the relevant domain. Invoking this "face validity" criterion (e.g., Anastasi, 1988) has always been inappropriate psychometrically for establishing the relations of items to a common underlying construct. For the psychometric models that we are proposing, we need to consider models that account for within-construct (intraconstruct) versus between-construct (interconstruct) order effects. We discuss intraconstruct effects first.

Intraconstruct Item-Order Effects

To illustrate how intraconstruct item-order effects can be evaluated, consider the abortion items from the GSS. Order effects have been reported for these items by juxtaposing single items that assess possible reasons for obtaining an abortion and finding that endorsement rates vary as a function of the order in which the items are juxtaposed. We shall assume in this particular discussion that each of the GSS items is hypothesized to be tapping a *single* unobserved abortion attitude. Under this assumption, one might assess the "strength" of a respondent's abortion attitude by summing across all presented abortion items or establish the reliability of such a measure by using an internal consistency measure such as the alpha coefficient.

Items may or may not be equally good indicators of a single abortion attitude. Such differences may be due to characteristics of the items (e.g., Hippler & Schwarz, 1987), characteristics of the person responding to the items (e.g., Judd & Lusk, 1984; Leone & Ensley, 1986; Tesser, 1978), or the interaction of person and item (e.g., Panter, 1989). If items do not tap this dimension to the same degree, then order effects will be confounded, as will be demonstrated in our consideration of psychometric models. Items may also be hypothesized to represent different components of abortion attitudes, but we shall treat this case in the next subsection on interconstruct order effects.

The logic of the intraconstruct approach to testing item order parallels the one made about subjects. There it is assumed (but again typically not tested) that, through random sampling mechanisms, individuals are interchangable units of analysis, and individual differences are ignored. In the case of items, it might be assumed (but again not tested) that the two presented items (i.e., the general item and the specific item) are sampled from the universe of all possible items that might be thought to measure a single underlying attitude toward abortion. Of course, this metaphor breaks down because items are generally not considered as random variables.[2] In attitudinal research, individual items are often used in research contexts on the basis of content alone, without regard to the specific properties and characteristics of the items themselves. Item selection from a larger pool of items is a process that requires both content considerations and empirical analysis, since either alone can be misleading (Jackson, 1971; Wainer & Braun, 1988).

Interconstruct Item-Order Effects

In contrast to the prior case where we assumed that items assess the same unidimensional construct, item-order effects can also be conceptualized as occurring across different constructs. This research emphasis typifies the situation in survey applications where practical considerations may preclude administering multiple items that assess a domain of research interest. For example, instead of assuming that the GSS abortion items tap a single dimension of attitudes toward abortion, we might assume for this case that each item assesses unique and multiple dimensions of abortion attitudes.

More concretely, the general and specific items would be hypothesized to measure distinct (and, ideally, independent) dimensions of abortion attitudes. Although there is clearly a conceptual relation among the GSS abortion items (i.e., they all ask respondents to express endorsement of situations under which a woman could obtain an abortion), this perspective underscores the distinct (as opposed to common) aspects of what these items are assessing. Given that these items are thought to tap distinct constructs, we can expect the internal consistency reliability among these items (as might be assessed by coefficient alpha) to be low.

The difference between intra- and interconstruct approaches might best be illustrated by example. Consider a sample of respondents who are presented with items concerning how much they like different films. The listed films are *Star Wars, The Wizard of Oz, The Sound bf Music, Lawrence of Arabia, Annie Hall, ET,* and *West Side Story*. If the obtained responses to these items across subjects reflected the single dimension of "liking movies," we would expect items to demonstrate strong internal consistency, reflecting their common dependence on "liking movies." On the other hand, the item responses might reflect a more complicated process such as a two-dimensional clustering (representing, for example, "liking musicals" and "liking nonmusicals"). In either case, we believe that whatever item associations are obtained in data reflect some implicit organizational structure that characterizes an average respondent's understanding and interpretation of the presented items.

From the perspective of the psychometric models to be discussed in this chapter, we can account for either an intra- or interconstruct representation for items. The choice between these representations is an investigator's prerogative and will be guided by theoretical concerns. However, given these distinctions, the choice between these alternatives must be specified and empirically validated. From the predominantly intuitive perspective that has characterized much of the item-order research to date, this choice has been arbitrated simply by assumption and has not always been stated and tested explicitly.

[2] Of course, this logic does not apply for random effects designs such as those assumed in generalizability theory (Brennan, 1983).

The use of items without specifying the hypothesized relations to underlying constructs is one aspect of this problem considered from a psychometric perspective. In addition to model specification, we believe that a more systematic understanding of item properties can be obtained through empirical tests. Such tests evaluate the equivalence of items as they relate to underlying construct(s). We shall draw on both classic and modern psychometric theories for these tests.

Psychometric Issues

In the methods that we discuss, latent variables describing unobservable processes are invoked. In test theories, latent variables have been described under the rubric of "true" scores. Before we proceed in outlining the statistical preconditions, we draw some distinctions about the implications of belief in true scores or latent variables.

Are True Scores "True"?

The assumptions that underlie the theory of "true" scores in measurement models have had their critics (e.g., Cliff, 1983; Knowles, 1988; Lumsden, 1976; Schönemann & Steiger, 1976; Strack, personal communication, October 1989). Among methodologists, the idea of "true" scores or latent variables has been attacked on the basis of the indeterminacy associated with relating indicators to latent variables and the lack of parsimony from introducing a higher level of abstraction than necessary for characterizing empirical relations. Substantive researchers claim that the psychometrician's conceptualization of "true" score differs from the subjects' phenomenological perspectives that they may (Ericsson & Simon, 1984) or may not (Nisbett & Wilson, 1977) be able to report on with accuracy.

From the psychometric and statistical viewpoint, we also feel some discomfort in postulating a classical test theory true-score model for measured variables. After all, it is tautological reasoning that decomposes an observed variable into two unobserved components (true score and error). From the perspective of psychologists interested in substantive issues, high premiums are generally placed on having high-quality behavioral data that are free of the potential biases of other methods of data collection (e.g., self-report). As empirically trained scientists, we have learned to value that which we can observe.

On the other hand, in much of contemporary psychology, interest is not in behavior per se but in behavior as an index of some underlying construct. The response to an attitude item is of interest only to the extent that it is a reliable indicator of the generating attitude structure within the individual. As we have argued elsewhere (Tanaka et al., 1990), most current psychological theories employ unobservable constructs (e.g., personality, attitudes, cognitive processes) that are operationalized in terms of specific observable measures.

The importance of the underlying construct is particularly true of current process-oriented models that dominate the literature on order effects. Psychometric theories can formalize and evaluate the links between observed measures and unobservable constructs. Thus, psychologists can move away from the idea of attitude assessment by assumption toward formal mathematical and statistical models and can evaluate the validity of their indicators.

"True" scores may or may not be true in any phenomenological sense. However, in the psychometrically based framework adopted here, formal testable models are considered preferable to implicit untestable models. The "true"-score concept as we use it in our discussions represents the association of measured variables as indices of unobservable constructs such as personality, attitudes, and cognitive processes. We next describe how effects can be empirically modeled when items are thought to tap a single dimension.

Intraconstruct Effects and Classical Test Theory

A major contribution to explaining the different kinds of relations that might exist between items and an underlying unidimensional construct was provided by Jöreskog (1974), who formalized a set of testable models of relations among measures of a construct. Specifically, a taxonomy of different models was described for how a group of items might relate to a hypothesized single underlying construct, such as an abortion attitude. In the Jöreskog hierarchy, each level implies increasingly stringent assumptions about relations between observed measures and the underlying construct.

The weakest of the model assumptions states that all itèms within a set are indicators of a single dimension. These items may be differentially related to the underlying construct; that is, some items may be more strongly related to the underlying construct, whereas others may show weaker relations, but all items are assumed to reflect the same underlying dimension. Such items are said to be "congeneric." Differences among cohgeneric items in terms of their relation to an underlying construct might be caused by wording, question length, or other such properties (e.g., Hippler & Schwarz, 1987). Thus, if the GSS abortion items are congeneric, they might have relations of different magnitudes to the underlying single abortion attitude, but they all would be reflections of that dimension.

A stronger assumption than congeneric measurement is that target items are related in a uniform way to the underlying attitude, in addition to relating to the same dimension. Thus, in this case, if one were able to compute the zero-order correlation between any particular item and the underlying construct, then these correlations would be identical for all items. Such measurements are called "tau-equivalent," following from the classic psychometric evaluation of "true" scores (tau) for a construct.

The most stringent version of the assumption regarding item equivalence is that, beyond a common homogeneous relation to a single underlying construct, all items are equally reliable. Thus, items would be related to the underlying construct in identical ways, and their errors of measurement would also be constant. From the classic psychometric perspective, groups of such items are referred to as "parallel" measures.

Jöreskog (1974) described a statistical framework to evaluate each of these assumptions that employs linear structural equation modeling techniques, which are well suited for testing process hypotheses (e.g., Tanaka et al., 1990). However, there is one problem in the direct adoption of the Jöreskog framework for testing item order in many measurement situations in the survey literature. The approach was intended initially for interval-level continuous measurement. In our example with the GSS items, the unit of analysis is the single item. As has been suggested by Muthén (1984) and others, the application of statistical methods for continuous data to what are essentially ordered categories or even dichotomies of item responses may lead to conclusions that are not correct. Thus, to apply the Jöreskog approach, we must look to developments in modeling of ordered categorical and/or dichotomous data and item response theory.

Factor analysis (cf. Harman, 1976) is the most well known method for evaluating the congeneric model. In factor analysis, a matrix of interitem correlations is modeled empirically to discover (in an exploratory framework) or test (in a confirmatory framework) the underlying dimensionality of constructs generating the responses to the measures. However, the dichotomous nature of the GSS items poses a problem for the traditional factor model. Research in the factor analysis of dichotomous items (e.g., Ferguson, 1941 [summarized in Comrey, 1973]) has suggested that the simple factor analysis of correlation coefficients for dichotomous items may lead to inappropriate conclusions, with a tendency to suggest more dimensions than necessary for explaining the intermeasure structure of the data. Fortunately, recent psychometric developments have presented alternative methods for examining the structure of dichotomous data.

One proposed approach is an extension of work by Bock and Lieberman (1970) and Bock and Aitken (1981). This approach is implemented in the computer program TESTFACT (D. Wilson, Wood, & Gibbons, 1984). An alternative approach (Muthén, 1978, 1987; Muthén & Christoffersson, 1981) also provides a method for factor analyzing dichotomous data and is implemented in the computer program LISCOMP (Muthén, 1987).

Some broad distinctions can be made between the approaches for dealing with noncontinuously measured data. For example, the LISCOMP model is flexible and can be adapted for exploratory or confirmatory factor analyses, whereas the TESTFACT model is appropriate only for exploratory factor analysis models. Moreover, the TESTFACT model is best suited for a large number of variables (and a small number of hypothesized factors), whereas LISCOMP is limited in the number of variables that it can analyze. Finally, extensions of the LISCOMP model to ordered categorical, truncated, and censored variables are quite easy, whereas TESTFACT is limited to dichotomous data. Further comparisons and contrasts between these two approaches can be found in Mislevy (1986).

Evidence against the Congeneric Assessment of the Abortion Items

In the research conducted to date on order effects in the GSS abortion items, it has been assumed that all items measure the single underlying construct "attitudes toward abortion." As such, responses to the juxtaposed GSS abortion

items are hypothesized to be cognitively influenced by the way in which subjects encounter, process, and judge these items on a single attitudinal structure.

In considering the factor structure of these dichotomous GSS abortion data, it is fortuitous that Muthén has employed this data set in multiple analyses to demonstrate his techniques for dichotomous and ordered categorical data (e.g., Muthén, 1978, 1981; Muthén & Christoffersson, 1981). In all of these analyses, he easily rejects the hypothesis of unidimensionality for these items, interpreting a two-factor solution with abortion for social reasons (e.g., the item asking whether a legal abortion should be possible "if she is married and doesn't want any more children") and abortion for medical reasons (e.g., the item asking whether a pregnant woman should be able to obtain a legal abortion "if there is a strong chance of serious defect in the baby") comprising the two factors. Interestingly, in some of the work on order and context effects with these items (e.g., Schuman et al., 1981), the two items that have been juxtaposed experimentally come from these two domains. To draw appropriate conclusions about order effects where an intraconstruct situation is assumed, it must be shown empirically that the same domain or knowledge structure is being assessed by all presented items. Yet, the Muthén results disconfirm the hypothesis that the item set is unidimensional.

Items Must Be Uniformly Related to the Underlying Construct

If the unidimensionality assumption can be met within a set of items for which an order effect has been demonstrated, other conditions also need to hold true if such effects are to be interpreted unambiguously. The previously discussed psychometric assumptions suggest other conditions could be tested. Because of its flexibility in testing some necessary constraints on the model, we frame our discussion in the context of Muthén's LISCOMP model. However, we acknowledge the models that we present here and those discussed in Jöreskog (1974) in classical test theory models. At the item level, our development focuses not only on the relations of items to their underlying construct but also on possible homogeneity of what have been termed "item difficulties." We believe that these points are best considered through another popular psychometric model, item response theory (IRT). We do not intend to provide an extensive review of item response theory here but refer interested readers to other sources (e.g., Hambleton & Swaminathan, 1985; F. M. Lord, 1980).

Developed in the context of educational testing, IRT relates the conditional probability of an observed item response to both item characteristics and person "ability." Although the "ability" interpretation of the person parameter has predominated in the IRT literature (given its use in educational domains), this parameter refers to any unidimensional latent construct, such as a trait or an attitude. In the context of attitude or personality items, the parameter may be conceptualized as an individual's proclivity to respond to items assessing the unidimensional construct (e.g., Reise & Waller, 1990; Reiser, 1980).

In one standard representation of the IRT model, the two-parameter model, two item parameters are hypothesized to relate to the conditional probability of an observed response given an individual's standing on the underlying attribute. Item discriminations refer to the strength of the relation between the observed response and the unobserved construct. Thus, item discriminations might be viewed as analogous to factor loadings in the more familiar factor analysis model.

Item difficulties characterize the tendency of respondents to endorse a particular item. Considered from the perspective of ability applications of the IRT model, items ordered on their difficulties might be indicative of increasing mastery of the domain. From the perspective of attitudinal or survey items, item difficulties might be conceptualized as representing the response extremity of presented items. For example, a set of Guttman-scaled attitude items would represent a set of items ordered in terms of their item difficulties from difficulties small in magnitude to those large in magnitude.

The hypothesis of item-order effects can be viewed as hypotheses about item difficulties. Thus, if cognitive processing is affected by item order, then effects should be observed at the level of item difficulties (i.e., the relative response proclivity to an item given a particular ordering). However, to establish such effects for item difficulties, it must first be demonstrated that the items relate to the underlying attitudinal construct in the same way across treatment conditions (e.g., equal item discriminations). Context should not change the relation of the item to the construct. We shall describe in the next subsection how this is accomplished using latent variable modeling procedures.

Although requiring equal discriminations might be viewed as rather stringent, it is essentially the same as the homoscedasticity assumption in the analysis of variance, in which hypotheses about mean differences between groups cannot be evaluated formally without assuming equal dispersion across those groups.

Assessing the Preconditions for Item-Order Effects

We next describe how to develop testable models of item-order effects. In discussing these ideas, we rely on the Jöreskog (1974) discussion of congeneric, tau-equivalent, and parallel models. These models can be easily adapted to an IRT perspective, in which the unit of analysis is the item. Thus, in this development, we take advantage of the unique properties of dichotomous or ordered categorical data.

Congeneric models reflect the least restrictive of these measurement models in the classical test theory framework for unidimensional phenomena. Discussions of item unidimensionality in the IRT framework are discussed, for example, by Reckase (1979, 1985). Previous work with, for example, the GSS abortion items would appear to suggest that they are not congeneric assessments.

Although we suggested that the GSS abortion items may not be congeneric, let us assume that a set of congeneric items could be identified and that the assumption of unidimensionality was found to be plausible. Two further assumptions would then have to be shown to hold true to allow unambiguous conclusions about item-order effects. First, following the tau-equivalence assumption, items would have to be shown to relate in a uniform way to the underlying construct, since context could not change the relation of the item to the underlying construct. Having met the congeneric and tau-equivalence preconditions, we could then test the null hypothesis of equal item difficulties or the assumption of parallelism. We elaborate each of these precondition tests in turn.

Tau-equivalence requires that measures are uniformly related to the underlying construct. Another way of stating this characterization is that the zero-order correlation between any measure (item) and the underlying construct is constant for all items. In the IRT framework, this is the assumption of equal item discriminations as found in the Rasch (1980) model. It is interesting to note that the sociologist Duncan (1984) has been a major proponent for the use of the Rasch model as an operating measurement model in social measurement. The tau-equivalence assumption, although quite stringent empirically, could be addressed in a number of ways. For example, items might be pretested, so that the zero-order relations between items and the construct were constant. Dichotomous response format items might be written in such a way to maximize variance; in other words, items would be worded so that endorsement rates would be approximately 0.5 for each response alternative. If observed data were consistent with the Rasch model, tau-equivalence could be established by testing the equality of item–total correlations. However, it is probably not safe to assume that any set of items will have properties such that the tau-equivalence assumption will be routinely met.

In the development of his model, Muthén suggests that latent response variables reflect underlying normally distributed variables that have been cut at some threshold to generate the observed dichotomous (or ordered categorical) process. This is the same logic that calculates tetrachoric correlations for dichotomous variables. These correlations form the data foundations of Muthén's model.

The tau-equivalence assumption is still only a necessary precondition for the interpretation of order effects. The most stringent hypothesis to be tested to establish that such order effects exist depends on a rejection of the null hypothesis of parallel assessments, given the existence of tau-equivalent assessments. To establish that item-order effects exist, it is necessary to operate under the assumption that items are equally good measures of the underlying construct, in terms of both their difficulty and their relations to the underlying construct or discriminability. If this is true, then the items are said to be parallel with no differences at either the mean or covariance level. We believe that it is only in this context that the issue of item-order effects can be addressed.

Items versus Item Clusters in the Examination of Context Effects

Despite the stringent evaluation of how items are related to an underlying attitudinal construct, information at the item level will generally be of limited utility. As indices of underlying attitudes (and all other things being equal), single items are more likely than multiple items or item clusters to be unreliable and to elicit inconsistent responses. Hence, a preferable method might be to aggregate (reliable) items within a particular question type (e.g., the general-type or specific-type questions) and then see whether effects currently being interpreted at the item level might also appear when looking at these aggregated item clusters. In the educational literature, a number of authors (e.g., Wainer & Kiely, 1987; Wainer & Lewis, 1990) have made similar observations to help clarify context effects involving ability items. We shall outline the logic of this aggregate approach or the use of "testlets," as they are referred to in the educational measurement literature.

Testlets are subsets of items, which are placed together and treated as units on the basis of some theoretical or content specifications. Wainer and Lewis (1990) review three examples of testlets. For example, all items that directly refer to a single reading-comprehension passage might be viewed as a testlet. Similarly, items that are hypothesized to be general might be clustered together with the assumption that their "generality" is their dominant, shared attribute. Such a "general" testlet might then be compared to a "specific" testlet constructed in an analogous manner. There are a number of reasons why item-order research might want to begin considering testlets instead of single items.

Conceptualizing items in terms of testlets or units of related items rather than single pieces of information changes item-order analysis. First, well-constructed testlets are comprised of items that are related to one another according to some content specification such as topical similarity or content balance and will necessarily be psychometrically more reliable units of analysis than will single items. Second, testlets can be evaluated in terms of the parallelism hypothesis that we previously discussed. Although this may be an unreasonably stringent assumption, parallelism of items within a testlet can provide researchers with relative security about (a) the properties of testlet items, testlets, and their underlying dimensions and (b) how items are perceived by respondents compared with other items within the testlet. Thus, as Wainer and Kiely (1987) note, "Each item is embedded in a predeveloped testlet, in effect carrying its own context with it" (p. 190). It might be desirable to create testlets whose components (items) have known relations to one another to control for item-rearrangement effects, particularly when certain items are deemed critical.

Finally, once testlets are constructed and treated as the unit of analysis, hypotheses based on context effects or the invariance of item and/or attribute parameters can be examined between testlets. The advantage in this case is that the tau-equivalence of the testlets (i.e., whether each testlet equally relates to the underlying construct) can be determined and tested in a straightforward manner.

Beyond the psychometric strengths that we perceive in employing the testlet strategy to investigate item-order effects, we feel that a claim for the superiority of this approach can also be based on theoretical and conceptual foundations. Consider, for example, the frequency-of-activation models in social cognition and social judgment reviewed by Wyer and Srull (1989). In these models, the frequency with which a construct is activated increases its accessibility. This increased accessibility, in turn, has demonstrated robust effects on social judgment and decision making. In the present context, the availability of multiple items in a testlet should repeatedly activate the construct that the items are presumed to measure. In moving from a general to a specific testlet (as opposed to single items), this increased activation level should predict a larger effect than has been previously obtained, given the results of these models.[3] Thus, we believe that testlets represent one way to begin interpreting these order effects with the confidence of reliability and unambiguous interpretation.

Interconstruct Item-Order Effects

In addition to the intraconstruct case, where items being compared are hypothesized to represent the same latent dimension, a theoretical model involving items from different (potentially related) constructs may be developed. In this interconstruct case, item-order hypotheses can be tested using extensions of the framework of assumptions described above, with some modification. We can show that although, by definition, the assumption of congeneric assessments would not apply in the interconstruct case, the increasingly stringent assumptions of tau-equivalence and parallel models can be evaluated in item-order investigations examining multiple constructs, and the concept of testlets can be incorporated.

Consider the simple case where responses to two items (or testlets), each tapping conceptually different constructs, are compared in one order versus another. Owing to the factor analytic developments of Muthén and others discussed previously, we need not make assumptions about the exact measurement scale of the items under consideration; they may be continuous, ordered categorical, censored, truncated, or dichotomous. In the interconstruct case, models can be tested by making explicit statements about item relations with their hypothesized construct in one condition versus the second condition. Two out of the three model assumptions described for the intraconstruct case are relevant with multiple constructs.

The assumption of congeneric measurement questions whether items reflect the same underlying construct. As noted, the GSS abortion items employed in our examples were shown to reject the congeneric assumption and to reflect two conceptually different constructs, representing subcomponents of abortion attitudes, Medical Reasons for Abortion versus Social Reasons for Abortion. To the extent that these constructs represent theoretically interesting distinctions, they can (and should) be treated as such. Thus, with failure to find support for a congeneric model for these items, the available evidence suggests future models should be tested in an interconstruct framework.

When items from two constructs are compared, the tau-equivalence assumption could test whether the strength of each item's relation to its underlying construct (the item discrimination) varies as a function of order condition. Note that this emphasis is not on whether each item relates to its underlying construct

[3] It must also be noted that, given a frequency-of-activation model for these effects, it is possible that the conditions under which these effects might be obtained would be delimited (e.g., Herr, Sherman, & Fazio, 1983).

to the same degree; instead, the investigator seeks to determine the extent to which these relations between observed variables and latent constructs are invariant across order conditions. If invariance as a function of condition cannot be demonstrated, we are again in a position where noncomparable elements are being contrasted. However, this noncomparability may be diagnostic in that across-group comparisons of differences among individual parameters can be inspected to determine what differences emerged in item order.

Finally, the assumption of parallelism focuses the equivalence of relations to underlying constructs across conditions, as well as equal error variances for each item with respect to its construct. By adding the constraints on error variances, this assumption becomes quite stringent. Although this ensures metric comparability in an across-condition comparison, the particular goals of an investigation might arbitrate whether such criteria are necessary.

A multiple-sample analysis of these assumptions permits evaluation of whether the patterns describing the relations in order AB for item A and item B is the same (or fairly close) to that describing order BA in a between-subjects design (cf. Jöreskog & Sörbom, 1989). Clearly, the comparison of models between treatment conditions changes as a function of the choice of assumptions to be tested (i.e., tau-equivalence assumption).

As we have suggested in the context of intraconstruct hypotheses, investigators can begin to define concepts such as "general" or "specific" in terms of groups of items, or testlets, hypothesized to be similar along these dimensions. Testlets can function as items, although their use carries the additional benefit of aggregated itemmetric responses.

An example of this interconstruct hypothesis in the personality domain is given by Osberg (1985). In this study, he considered variations in responses to the Self-Consciousness Scale (SCS) of Fenigstein, Scheier, and Buss (1975) as a function of where this measure was placed in a group administration of this and other personality measures in a large sample. In the between-subjects design that he employed, Osberg administered the SCS in different serial positions along with other personality inventories (e.g., Locus of Control, Need for Cognition) to test hypotheses about level differences as a function of order. From the perspectives of the psychometric models proposed here, level differences in the SCS would be maximally relevant if it could be shown that (a) the SCS's psychometric properties do not change as a function of the order of presentation and (b) the tests in which the SCS is embedded (e.g., Locus of Control, Need for Cognition) also do not change in terms of their psychometric properties.

Concretely, these conditions would imply a model where items or testlets were linked to the SCS construct in each different order condition under the tau-equivalence or parallelism assumptions. Similarly, the items or testlets of the other personality measures would be linked to their respective constructs. The interconstruct hypothesis of level differences would be evaluated at the construct level. This hypothesis would address the issue of whether, at the construct level, differences could be observed.

It is interesting to note in passing that current research in the interconstruct tradition is implicitly assuming a Rasch measurement model for items assessing the constructs. In a simple two-construct case, this operating model assumes equal links between items and constructs for each of the two (unidimensional) constructs. Further, individuals' total scores accurately reflect their position on the distribution of the unobserved construct. It is in this case only that a between-group comparison of total scores would be psychometrically appropriate.

Summary

There are strong parallels between the sequence of research issues for evaluating order effects in the educational and ability literatures and issues that we have identified in the survey literature. In their review, Leary and Dorans (1985) identified three major stages of research on item-order effects. During the first stage, the phenomenon of item order was described and might be considered analogous to effects in the tradition of Schuman et al. (1981). The second phase of research attempted to identify processes responsible for these effects. In the survey literature, this is paralleled by the substantive contributions in this volume and its predecessor (Hippler et al., 1987) and the work of investigators modeling the survey response process using principles in cognitive and sociocognitive psychology. The focus of this chapter parallels the third stage identified by Leary and Dorans. In this stage, item-order effects are evaluated only after the psychometrics of the items has been considered.

In this chapter, we presented the strong conditions under which item-order effects can be interpreted. We outlined some preconditions that, from a psychometric standpoint, must hold true to allow for interpretation of order effects. Differences were established between the intra- and interconstruct cases. The intraconstruct case, where items are thought to be tapping the same underlying construct, involves three testable models of increasing stringency. In the interconstruct case, only two of these models are applicable. Finally, we introduced the notion that researchers might begin to concentrate on item clusters or testlets to test for order effects. The use of testlets will serve to increase the reliability of obtained effects by creating a better controlled within-testlet context.

We recognize that it is unlikely that researchers will always operate under these stringent conditions. We hope that we have raised issues for investigators in this literature to consider. In designing future research on item-order effects, an awareness of the psychometric issues involved in item (or testlet) comparisons is crucial, and these issues must be considered before order effects can be interpreted.

Acknowledgments. We wish express our thanks to the conference organizers, Norbert Schwarz and Seymour Sudman, for allowing us to participate in the exciting intellectual exchange at the conference and to contribute to this volume, as well as for provid-